IEEE
NOVEMBER 2024, VOL. 7, NO. 6

IEEE INTERNET OF THINGS MAGAZINE

IoT Security and Provisioning in Cyber-Enabled Niche Critical Infrastructure

IEEE Internet of Things
A Publication of the IEEE Internet of Things Initiative, a Multi-Society Technical Group
www.iot.ieee.org

Publish your work

in the *IEEE Open Journal of the Communications Society* (OJ-COMS), the premier open access journal in communications technology!

The *IEEE Open Journal of the Communications Society* (OJ-COMS) is a fully open-access, all-electronic journal that publishes original high-quality, peer-reviewed manuscripts on advances in telecommunications systems and networks.

OJ-COMS reaches more than **5 million people** and offers a rapid review process. Submissions reporting new theoretical findings and practical contributions are welcome. Additionally, review and survey articles are considered.

Submit your work today!

www.comsoc.org/ojcoms

IEEE ComSoc | IEEE

IEEE
INTERNET OF THINGS MAGAZINE

NOVEMBER 2024, VOL. 7, NO. 6

IOT SECURITY AND PROVISIONING IN CYBER-ENABLED NICHE CRITICAL INFRASTRUCTURE

CONNECTED IOT SERVICES

CONNECTED VEHICLES

IEEE INTERNET OF THINGS MAGAZINE (ISSN 2576-3180) is published bimonthly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters address: IEEE, 3 Park Avenue, 17th Floor, New York, NY 10016-5997, USA; tel: +1 (212) 705-8900; https://www.comsoc.org/publications/magazines/ieee-internet-things-magazine. Responsibility for the contents rests upon authors of signed articles and not the IEEE or its members. Unless otherwise specified, the IEEE neither endorses nor sanctions any positions or actions espoused in IEEE Internet of Things Magazine.

ANNUAL SUBSCRIPTION: US$53 print subscription; US$38 electronic subscription; US$624 non-member print subscription.

EDITORIAL CORRESPONDENCE: Editor-in-Chief, Rath Vannithamby, rath.vannithamby@intel.com; Associate Editor-in-Chief, Maurice Khabbaz, maurice.khabbaz@gmail.com.

POSTMASTER: Send address changes to IEEE Internet of Things Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331. GST Registration No. 125634188. Printed in USA. Periodicals postage paid at New York, NY and at additional mailing offices. Canadian Post International Publications Mail (Canadian Distribution) Sales Agreement No. 40030962. Return undeliverable Canadian addresses to: Frontier, PO Box 1051, 1031 Helena Street, Fort Eire, ON L2A 6C7.

SUBSCRIPTIONS: Orders, address changes – IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08855-1331, USA; tel: +1 (732) 981-0060; e-mail: address.change@ieee.org.

ADVERTISING: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, IEEE Internet of Things Magazine, IEEE, 3 Park Avenue, 17th Floor, New York, NY 10016-5997.

SUBMISSIONS: The magazine welcomes tutorial or survey articles that span the breadth of IoT technologies and applications. Submissions will normally be approximately 4500 words, with no or few mathematical formulas, accompanied by up to six figures and/or tables, with up to 15 carefully selected references. Electronic submissions are preferred and should be submitted through Manuscript Central: https://mc.manuscriptcentral.com/iotmag. All submissions will be peer reviewed. Submission instructions can be found at the following:

https://www.comsoc.org/publications/magazines/ieee-internet-things-magazine/article-submission-guidelines.

ACCEPTED FROM OPEN CALL

FUNDAMENTALS OF WI-FI

FROM **802.11AX** TO **802.11BE** AND BEYOND

7–8 November 2024
9:00 am–3:00 pm EST

Learn about Wi-Fi's evolution, focusing on key features of Wi-Fi 6/6E, Wi-Fi 7, and emerging developments in Wi-Fi 8, along with exploring its essential features, operations, and the exciting developments on the horizon.

To learn more or to register, scan the QR code or go to: https://comsoc.co/4es6vFx

IEEE COMSOC TRAINING

Online Course ▪ https://comsoc.co/4es6vFx

Mentor's Musings on Systems Thinking & Standardization Imperatives for Making the Critical Infrastructure Cyber Secure & Resilient

N. Kishor Narang
Technology Philanthropist, Ethicist, Innovation Standardization & Sustainability Evangelist

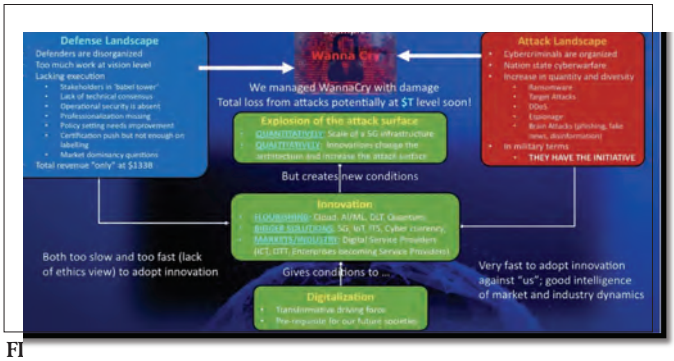
Introduction

Protecting the cybersecurity of critical infrastructures and their supply chains is crucial for the simple reason that these systems power our daily lives—from electricity and water to healthcare and transportation. A cyber incident disrupting the functioning of these vital services can cause widespread chaos, endanger lives, and cripple economies. As cyber threats grow increasingly sophisticated and pervasive, ensuring the resilience and security of these critical systems is not just a technological necessity but a fundamental safeguard for the well-being and continuity of modern life. It needs a Systems Thinking and System Engineering approach duly supported by comprehensive set appropriate Standards.

International law defines Four Global Commons (natural assets outside national jurisdiction) which are the earth's natural resources i.e. the High Seas, the Atmosphere, Antarctica, and Outer Space. Cyberspace is the 5th Global Common. It is also considered as the 5th Dimension beyond the 3 dimensions of Space & 4th dimension being the Time.

Challenges that all economies are facing today in safeguarding the security and privacy of its ecosystem including citizen are — Transnational Nature of Cyber Crime, "Cultural" Vulnerabilities, Internet Resilience and Threat Landscape.

Cyber risk threat vectors have evolved rapidly, and attacks have become increasingly sophisticated, deliberate, and unrelenting in nature. In the digital era, trust is a complex issue fraught with myriad existential threats to the enterprise. And while disruptive technologies are often viewed as vehicles for exponential growth, tech alone can't build long-term trust. Every aspect of an organization disrupted by technology represents an opportunity to gain or lose stakeholders' trust. Leaders are approaching trust not as a compliance issue but as a business-critical goal. For this reason, leading organizations are taking a 360-degree approach to maintain the high level of trust their stakeholders expect.



FI

The new paradigm of Smart Grid, Smart Home, Smart Building, Smart Manufacturing, Smart City already complicated by the 'Internet of Things' & Internet of 'Everything' made further complex by the 5G, Artificial Intelligence, Machine Learning, Blockchain & Quantum Computing, make it truly complex to develop and embed comprehensive Security, Privacy and Trustworthiness attributes in the products, systems and solutions for any use case or application — be it consumer, commercial, industrial, automotive or strategic domains like civic infrastructure.

The recent evolution of disruptive technologies and digitalization compounded by the Covid 19, changing geopolitical situations, and increasing cyber-attacks bring a whole new set of challenges for the Security and Security Evaluation Methodologies for complex nature & architectures of Civic and Citical Infrastructures of the nation leveraging the IT & Communication Networks evolving to meet these rising needs of the Society. While Artificial Intelligence (AI) (and Machine Learning (ML) is one of the technologies being deployed to mitigate cybersecurity risks due to its speed that human operators may not be able to match, not to forget that it is also being heavily leveraged for the cyber offensive endeavours to out maneouvre the cyber defenses of various digital systems/infrastructures. Hence it would be imperative to develop some comprehensive strategies to leverage AI & ML...

The highly protected Networks for the 'Critical Infrastructures' need to give access to the consumers for Consumer Engagement and Participation in these Smart (Digital) Infrastructures to meet the true drivers of setting them up. These large Smart Networks are actually highly complex 'Systems of Systems' and "Networks of Networks', and thus create fresh challenges in the Security Paradigm and development of Protection Profiles.

It is evident that Cyber Security is a very complex paradigm, and with evolving new technologies, requirements, and ever-increasing Attack Surface the vulnerabilities are rising many folds with time. In such a dynamic scenario, it is required to develop

a Cyber Security Strategy to make our Critical Infrastructure comprehensively Safe, Secure, Resilient and Trustworthy.

Digital Transformation

New technologies and paradigms like IoT, Big Data, Artificial Intelligence, Virtualization, Cloud Computing, 5G (& now 6G in the very near future), Blockchain, Quantum Computing and other disruptive technologies like AR/VR, Web 3.0 & Metaverse are promising to disrupt the way we design products, systems, and solutions. Today, we need to develop new strategies (and standards) that can help us navigate seamlessly through a much wider and complex canvas of technologies, ecosystem & stakeholders.

In the wake of technological advancements, especially in the field of ICT all the ecosystems be it Smart Grid, Smart Buildings, Smart Cities or Smart Factories now find themselves making three classes of transformations — Improvement of Infrastructure — to make it resilient & sustainable; Addition of the Digital Layer — which is the essence of the smart paradigm; and Business Process Transformation — necessary to capitalize on the investments in smart technologies. The genesis of Digital Transformation in any paradigm, domain or ecosystem being — "Sustainability is the True Destination"; "Resilience is the Core Characteristic'; 'Smart is merely the Accelerator"; and "Standards are the Chromosomes of Digital Infrastructure."

FIGURE 2. Digital Transformation constituents.

And, the Digital Transformation is not only about the integration of IT (Information Technologies) and OT (Operational Technologies) anymore rather a more complex confluence of cutting edge and disruptive technologies. The Network Technologies have also evolved to a level of complexity and advancements (5G & Software defined Networks) that play a crucial role in the digital transformation and hence need to be understood comprehensively beyond properly contextualizing as one of the core transformation constituent. Last but not the least, the new kids on the block – IoT & AI have become so pervasive and ubiquitous that they enhance the value proposition of Digital Transformation comprehensively in any ecosystem, and hence can't be overlooked when developing the Digital Transformation Strategy in any domain and/or ecosystem. It would be appropriate to consider that Digital Transformation is NOT a technology, rather it's a complex paradigm with domain-specific implications, as we are living in an ephemeral world, and Artificial Intelligence (AI) & Machine Learning (ML) and other disruptive technologies are powering the digital transformations happening in every domain around the world.

IoT & Critical Infrastructure Interplay

IoT refers to a network of interconnected devices, sensors, and objects that collect and exchange data. These devices can be anything from smartphones and wearables to smart home appliances, industrial sensors, or autonomous vehicles. With an unprecedented increase in the number of Internet of things (IoT) devices and emerging applications, a large amount of traffic is created every day. Such an increase poses a great burden on the Internet network and also demands significant investments for the infrastructure upgrade. However, thanks to the development of big data analytics and artificial intelligence (AI) techniques such as deep learning and machine learning, the data collected can be effectively exploited for many purposes.

Out of all the domains that IoT has revolutionized, critical infrastructure stands apart as a particularly tangible intersection of the digital and physical worlds. Today, it enhances operational efficiency, reduces costs, and steps up service reliability of electrical grids, municipal utilities, transportation systems, manufacturing entities, military facilities, airports, and more. This technological leap, predictably enough, comes with its challenges. First, deploying seamless IoT networks over long distances can require hefty engineering, construction, and investment to upgrade the existing wiring infrastructure or even build it from the ground up. Second, operating such networks is a tightrope to walk in terms of security, given the high-stakes assets at the heart of them.

The Cybersecurity Achilles Heel

The vast number of interconnected devices in an IoT-driven infrastructure creates a massive attack surface. These objects often have limited processing power and may miss out on robust security features, which potentially makes them easy targets. Here's a closer look at the specific concerns that shape up the unique threat model of an ecosystem like that:

- **Unauthorized access:** Many IoT devices have notoriously weak authentication protocols and are shipped with easy-to-guess default passwords that network administrators neglect to change. This leaves them vulnerable to brute-force attacks or credential stuffing.
- **Data breaches:** Without strong encryption in place, sensitive data transmitted between devices and control centers can be intercepted and mishandled.
- **Denial of Service (DoS):** IoT networks can be swamped by malformed queries whose number exceeds the server's processing capacity. This can result in significant downtime and operational issues that end up disrupting critical services.
- **Software vulnerabilities:** Outdated firmware and software on these devices can harbor unpatched security gaps, creating entry points for cyberattacks.

The catch-all thing to understand is that the very nature of interconnectedness creates vulnerabilities. Perpetrators targeting a single device could gain access to a wider network, potentially causing widespread disruptions.

IoT Security Done Right

As cyber threats evolve, overconfidence in defenses at the network perimeter can be a losing strategy. Even with top-notch proactive security measures in place, there's always a chance of well-motivated adversaries breaking in. It's best to prevent them from weaponizing the data they might intercept, in the first place. Call it a plan B, if you will, but it eventually pays off in today's nuanced cyberspace.

A good security philosophy would combine three layers of protection: end-to-end data encryption with MACsec 256-bit cryptographic standard, data fragmentation, and scrambling.

This means that to cause damage, a malicious actor would need to amass information from all nodes on the network in order to de-scramble it, put the fragments together in the correct order, and decrypt the resulting data with a unique key to make it meaningful. Anyone even remotely familiar with cryptography knows that this mission is close to impossible.

With the well-thought-out security approach (dubbed the Triple Shield) and a breakthrough hybrid-fiber network deployment principle, combined with NIST certification for FIPS 140-2 cryptographic standard, can predictably pave the way towards comprehensively cyber secure and resilient projects. Robust security requires more than just strong encryption. The Triple Shield approach integrates end-to-end data encryption, data fragmentation, and scrambling to create a multi-layered defense system. This ensures that even if one layer is compromised, the data remains protected through additional layers of security. By combining these techniques, it becomes exceptionally difficult for malicious actors to access and exploit sensitive information, maintaining the integrity and confidentiality of our clients' critical infrastructure.

What's Next for IoT-Driven Critical Infrastructure?

While security is crucial for networks that underlie critical infrastructure, enabling uninterrupted connectivity between IoT devices is another nontrivial challenge. This is especially true of geographically scattered environments that combine fiber, coax, and legacy copper wiring. The silver lining is that such heterogeneous cabling architectures can be glued together to deliver fiber-grade connectivity without the need to build new high-cost networks from scratch. Hybrid-fiber networking concept includes sections of fiber (for the easy-to-reach-with-fiber locations) and copper/coax that can be upgraded to run fiber-grade communication.

As the IoT element becomes instrumental in modernizing critical infrastructure across multiple industries, innovative network design principles come to the fore. The key challenge here is to avoid a tradeoff between deployment speed, ease of maintenance, and security. A safe world without serious technology-borne societal repercussions seems to be a matter of striking that balance for the long haul.

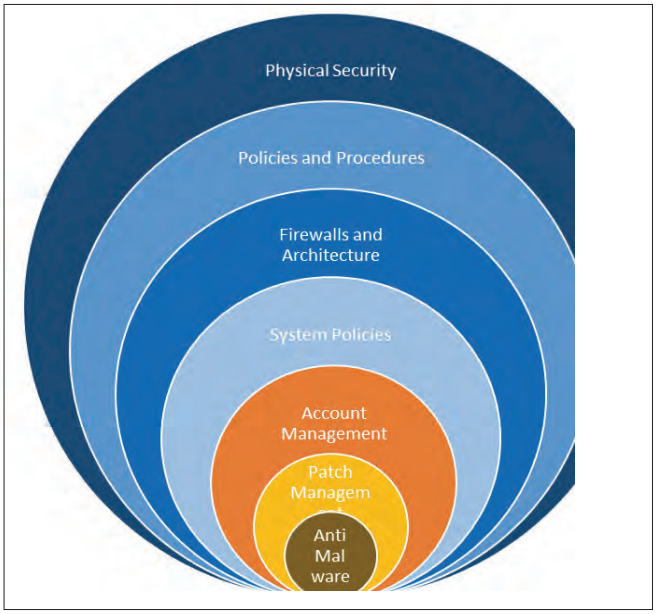


FIGURE 3. Defence in-depth approach for cyber security.

The Standardization Conundrum

The statement "**The beauty of Standards is, that there are so many to choose from...**" from Andrew S. Tannenbalm in 1990, is most apt for current situation in Security Standards. In last three decades of ICT evolution, an explosion in the standards-making ecosystem itself has further compounded the standards landscape. There are plenty of standards, and yet, we continue to identify more & more security standards gap in different ecosystems, use cases and applications. If we look at them without prejudice, we shall find that most of the security standards are Point solutions to Point Problems (Security Concerns). This is a typical bottoms up approach, which has over the time proven it to be quite Sub-optimal...

FIG

"The irony is that Standards & even SDOs are not at the forefront of Solution designers, developers, providers, deployers or users' minds." There are misconceptions on what standards are for, and the case for use of standards has not been made. Most researchers, design engineers and even start-ups argue that standards block innovation. In fact, Standardization brings innovation and spreads knowledge. Standardization helps define the contours of structured innovation, first because it provides structured methods and reliable data that save time in the innovation process and, second, because it makes it easier to disseminate ground-breaking ideas and knowledge about leading edge techniques. Liberalization and Markets have a lot of great virtues, but they cannot create their own conditions of existences: they must be designed!

The IoT value chain is perhaps the most diverse and complicated value chain of any industry or consortium that exists in the world. In fact, the gold rush to IoT is so pervasive that if you combine much of the value chains of most industry trade associations, standards bodies, the ecosystem partners of trade associations and standards bodies, and then add in the different technology providers feeding those industries, you get close to understanding the scope of the task. In this absolutely heterogeneous scenario, coming up with common harmonized standards is a major hurdle.

Most of the standards activity in the domain to date have been on the development of Communication Security, Device Security and IT Cyber Security standards that address individual limited Security concerns of different stakeholders. In case of Critical Infrastructure security, it is unlikely to be which standard, rather which standards since most architectures do not pick one standard but have a layered approach capable of using multiple standards in the portfolio.

Systems Approach

The multiplicity of technologies and their convergence in many new and emerging markets, however, particularly those involving large-scale infrastructure demand a top-down approach to standardization starting at the system or system-architecture rather than at the product level. Therefore, the systemic approach in standardization work can define and strengthen the systems approach throughout the technical community to ensure that highly complex market sectors can be proper-

ly addressed and supported. It promotes an increased co-operation with many other standards-developing organizations and relevant non-standards bodies needed on an international level. Further, standardization needs to be inclusive, top down and bottom up; a new hybrid model with a comprehensive approach is needed.

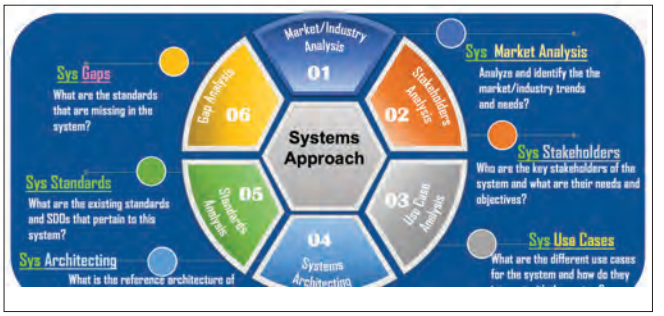
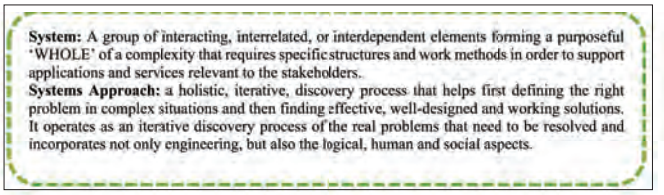


FIGURE 5. Systems approach process flow.

In fact, Systems Approach is about "Holism" inspired from the statement of great Philospopher Aristotle in 300 B.C. – "**The Whole is Greater than the Sum of its Parts.**"



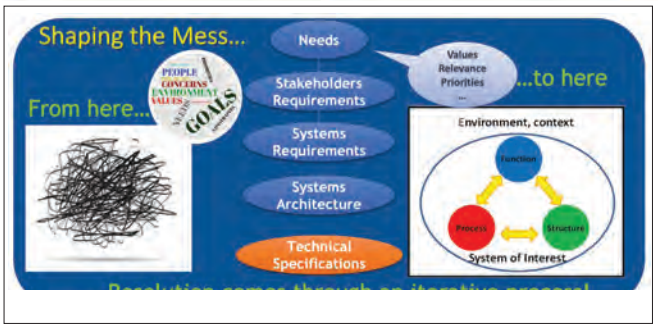


FIGURE 6. Shaping the mess…

Systems Approach Activities

- Identify and understand the relationships between the potential problems and opportunities in a real world situation.
- Gain a thorough understanding of the problem and describe a selected problem or opportunity in the context of its wider system and its environment.
- Synthesize viable system solutions to a selected problem or opportunity situation.
- Analyze and trade off between alternative solutions for a given time/cost/quality version of the problem.
- Measure and provide evidence of correct implementation and integration.
- Deploy, sustain, and apply a solution to help solve the problem (or exploit the opportunity).
- All of the above are considered within a life cycle framework which may need concurrent, recursive and iterative applications of some or all of the systems approach.

It is evident that Cyber Security is a very complex paradigm, and with evolving new technologies, requirements, and ever-increasing Attack Surface the vulnerabilities are rising many folds

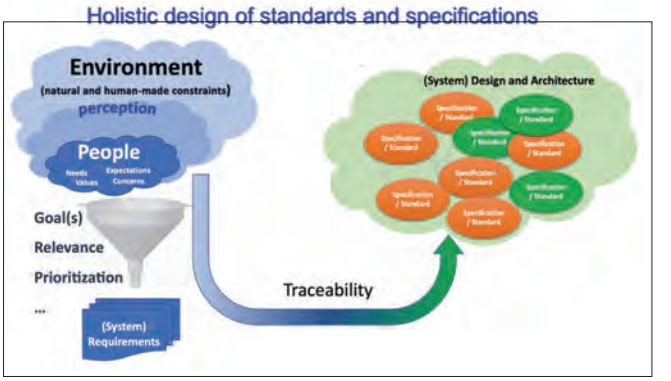


FIGURE 7. Holistic design of standards & specifications.

with time. In such a dynamic scenario, it is required to develop an equally dynamic & scalable Cyber Security Standardization Strategy and Roadmap to make our Digital Infrastructures **comprehensively** Safe, Secure, Resilient and Trustworthy.

Conclusion

We are entering an era of a cyber arms race where AI will amplify the impact for both the security professional and the adversary. Organizations cannot afford to fall behind, and the legacy technology of yesterday is no match for the speed and sophistication of the modern adversary. The speed and ferocity of cyberattacks continue to accelerate as adversaries compress the time between initial entry, lateral movement and breach. At the same time, the rise of generative AI has the potential to lower the barrier of entry for low-skilled adversaries, making it easier to launch attacks that are more sophisticated and state of the art. These trends are driving a tectonic shift in the security landscape and the world. The "good enough" approach to cybersecurity is simply no longer good enough for modern threats.

Cyber Immunity & Cyber Resilience

- The pandemic-induced digital transformation has increased exposure to cyber threats as we cross the digital fault line due to remote working and escalated online presence. To counter this, an intuitive and adaptive cyber posture defined by zero latency networks and quantum leaps will be needed across industries. These developments, while great for humanity, will challenge privilege, privacy, and defend every citizen.
- The speed of processing of AI systems is currently seen as providing protection for infrastructures and networks that human operators may not be able to match, especially as cyber-attackers are employing increasingly sophisticated methodologies. AI can potentially respond to a cyberattack scenario far more quickly than a human decision maker.
- Cyber Immunity at every layer will create networks that are inherently secure and self-learning. AI-induced digital intuition is one of the pillars of cyber-Security strategy that will allow intelligent adaption. **The ability of AI systems to out-innovate malicious attacks by mimicking various aspects of human immunity will be the line of defence** to attain cyber resilience based on both supervised and unsupervised machine learning.
- These systems will be designed to make the right decisions with the context-based data, pre-empt attacks on the basis of initial indicators of compromise or attack, and take intuitive remediated measures, allowing any digital infrastructure and organization to be more Resilient.

Build a cybersecurity culture — Though technology is clearly critical in the fight to detect and stop intrusions, the end user remains a crucial link in the chain to stop breaches. User aware-

ness programs should be initiated to combat the continued threat of phishing and related social engineering techniques. For security teams, practice makes perfect. Encourage an environment that routinely performs tabletop exercises and red/blue teaming to identify gaps and eliminate weaknesses in your cybersecurity practices and response.

Biography

N. Kishor Narang (kishor@narnix.com) is Technology Advisor, Mentor, Design Strategist & Architect in Electrical, Electronics & ICT with over 47 years of professional experience in education, research, design and advisory running an Independent Design House — NARNIX since 1981. Over 37 years of hardcore Research and Design Development Experience in Solutions, Systems, Products, Hardware, Software & Firmware (Embedded Software) across diverse technology & application domains, and over 10 years of Strategic Advisory Experience to different segments of business & industry. He has over 500 Research & Design Mentees in the Electronics, ICT & STI (Science Technology & Innovation) Ecosystems. Leading multiple National & Global Standardization Initiatives & Projects at BIS, TSDSI, IEEE, IEC, ISO, ITU and IETF... For the last 15 years, been deeply involved in standardization in the electrical, electronics, communications, information technology, digital infrastructure and cyber security domains with a focus on identifying gaps in standards to bring harmonization through system standards and standardized interfaces to ensure end-to-end Interoperability. Recent work includes advocacy and standardization in cross cutting societal Ethical imperatives — Sustainability (focus on Decarbonization in every domain), Safety of the Mankind and the Environment in use of disruptive digital technologies (including but NOT limited to Artificial Intelligence). Mentoring many Deep Tech & Disruptive Tech Startups in the domains of e-mobility, drones, robotics, automotive like Electric Vehicles, Autonomous Vehicles, Drones, Robotics & AI; Strategic Electronics like Defense, Aerospace, AR/VR/XR etc...

IEEE ComSoc Publications

IEEE Communications Society (ComSoc) publications deliver timely, in-depth, technical information on a wide array of communications technology topics that directly impact business and advance research for the benefit of humanity.

Magazines

ComSoc's award-winning, peer-reviewed magazines cover the latest issues and advances in key areas such as wireless communications, standards, and global internetworking.

Journals

ComSoc's journals earn the highest marks by the Journal Citation Reports® (JCR) and include high-quality manuscripts covering state-of-the-art research in a variety of wireless and telecommunications topics.

More Offerings:

Best Readings | CTN | GCN | Pubs Digest | TCN | Tech Focus | White Papers

www.comsoc.org/publications

IEEE ComSoc®

IoT Security and Provisioning in Cyber-Enabled Niche Critical Infrastructure


Elias Bou-Harb


Edgar Weippl


Chadi Assi


Martin Husák


Jiangshan Yu


Katherine A. Flanigan

This editorial introduces the Special Issue (SI) of *IEEE IoT Magazine*, focusing on IoT Security and Provisioning in Cyber-Enabled Niche Critical Infrastructure (CI). As digital transformation reshapes industries worldwide, CI sectors such as water and wastewater systems, manufacturing plants, chemical facilities, maritime operations, and the electrification of transportation fleets are at the forefront of this evolution. These sectors, once heavily reliant on isolated, physical systems, are now being reimagined through the lens of the Internet of Things (IoT). Sophisticated IoT-enabled sensors and devices are being integrated into core operations, offering unprecedented opportunities for real-time monitoring, automation, and data-driven decision-making. The rationale behind this shift is clear: increased efficiency, more precise control over physical processes, and the ability to extract actionable insights from vast streams of operational data.

However, these advances come with significant security challenges. As these previously siloed systems become interconnected and cyber-enabled, they also become more vulnerable to a variety of cyber threats. This growing exposure has already been exploited in real-world incidents, where both state-sponsored and independent malicious actors have targeted IoT vulnerabilities in CI. The consequences of such attacks are dire, extending beyond operational disruptions to impact national security, economic stability, and even human lives. Given the critical role these sectors play in society, it is imperative to develop robust strategies for securing these IoT-driven infrastructures. The motivation behind this Special Issue is hence to address this urgent need by providing a platform for research that tackles the unique security and provisioning challenges posed by niche critical infrastructure sectors.

In an era where cyber-physical systems are becoming the backbone of critical infrastructure, the rationale for this SI lies in its focus on sectors that are often overlooked but are nonetheless essential to national well-being. These niche infrastructure sectors, due to their unique technological setups and deployment scenarios, face distinct challenges that require dedicated research attention. Through this Special Issue, we aim to spotlight these challenges, inspire innovative solutions, and foster the development of practical applications that will secure these systems in the face of evolving threats.

We are excited to shed the light on six articles featured in this SI, which present a diverse array of research findings that delve into both technical and socio-technical aspects of IoT security in critical infrastructure. These contributions explore the identification of IoT-centric vulnerabilities, empirical analysis of system artifacts, novel attack modeling techniques, and innovative defense mechanisms. Beyond the technical aspects, the articles also address human factors, regulatory frameworks, and risk management strategies — elements that are equally critical for the resilience and safety of these vital systems.

Empirical Data Analysis

In the article "The Role of Rule Mining in Generating Synthetic Cyber-Physical System Attack Samples," the authors present a novel approach to addressing the critical challenge of creating realistic attack data for cyber-physical systems (CPS) by leveraging rule mining techniques. The generation of synthetic datasets through frequent itemset mining is central to their method, allowing for the simulation of varied and complex cyber-attacks in industrial control systems (ICS) where real-world attack data is often scarce or incomplete. By creating synthetic samples that capture the operational dynamics of CPS environments while incorporating controlled randomness, the article demonstrates how these datasets can enhance the development and testing of machine learning models for ICS security. This methodology overcomes the limitations of data availability, creating balanced datasets that can train more accurate and generalized defense mechanisms against a broad spectrum of cyber-attacks. The alignment with the MITRE ATT&CK framework further illustrates the potential of these synthetic samples to model real-world adversarial behavior in ICS environments, thus advancing the reliability of cyber security solutions in industries like energy, water, and manufacturing. Looking forward, the integration of rule mining with more advanced machine learning techniques holds promise for generating even more realistic datasets, enabling the continuous evolution of cyber security defenses in response to emerging threats. This research thus makes a substantial contribution to the field, laying the groundwork for future innovations in safeguarding critical infrastructure.

Intelligent Detection and Response

The article "Toward Intelligent IoT Endpoint Detection and Response Using Digital Twins via Firmware Emulation," highlights the pressing need for advanced cyber security measures as the proliferation of IoT devices introduces a multitude of

vulnerabilities across various industries. This work is significant as it presents a novel framework that leverages the concepts of digital twins and firmware emulation to enhance the detection and response capabilities for IoT endpoints, effectively addressing the critical security risks posed by these interconnected devices. The approach is particularly impactful because it provides organizations with a comprehensive methodology to proactively identify and mitigate potential threats, thus safeguarding sensitive data and maintaining operational continuity. By simulating the behavior of IoT devices through digital twins, the proposed framework enables real-time monitoring and analysis of device behavior, allowing for the detection of anomalies indicative of potential security breaches. Moreover, the practical usefulness of this research is underscored by its applicability across various sectors, including smart homes, healthcare, and industrial automation, where the security of IoT endpoints is paramount. The integration of firmware emulation further enhances the system's capabilities, allowing for a deeper understanding of device vulnerabilities and facilitating more effective threat response strategies. Looking ahead, future directions could involve the incorporation of machine learning algorithms to improve threat detection accuracy and the development of more extensive datasets for training purposes. Additionally, researchers could explore how these techniques can be adapted to the evolving threat landscape, including emerging IoT standards and protocols, thereby strengthening the overall security posture of IoT environments. By advancing the methodologies for IoT endpoint detection and response, this article lays the groundwork for future innovations in cyber security, ultimately contributing to a safer and more resilient digital ecosystem.

Integrated Threat Defense

The authors of "Merging Threat Modeling with Threat Hunting for Dynamic Cybersecurity Defense" represents a transformative shift toward a more proactive defense mechanism. This article articulates the pressing necessity for organizations to adopt an integrated approach, thereby enhancing their capacity to detect and respond to attacks in real-time. By synthesizing these two methodologies, the authors not only present a robust framework for improved attack detection but also elucidate how organizations can fortify their overall security posture. Continuous updates to threat models — derived from real-time intelligence, both internal and external — ensure that organizations remain vigilant against the ever-evolving threat landscape. The practical implications of this research are substantial. By outlining a structured integration of threat hunting and modeling, the article provides organizations with actionable strategies to optimize their cyber security operations. The introduction of automated tools and AI-driven analytics serves as a catalyst for organizations to better allocate resources, prioritize critical threats, and enhance cross-disciplinary training. This comprehensive approach not only streamlines operational efficiency but also empowers cyber security teams to develop the diverse skill sets necessary to navigate the complexities of contemporary cyber threats. Furthermore, the emphasis on continuous data stream integration lays the groundwork for a resilient security architecture that evolves in tandem with emerging threats. However, the journey towards effective integration is not without challenges. The article identifies several key areas for future exploration, including the need for automated integration techniques that merge threat hunting and modeling seamlessly, without compromising efficiency. The demand for advanced AI-driven models capable of real-time threat analysis is paramount, as organizations seek robust mechanisms to counteract stealthy attacks. Dynamic threat modeling also emerges as a critical research avenue, enabling real-time updates that leverage insights from threat hunting to keep threat models relevant and accurate. Integrating crowdsourced threat intelligence can further enrich these models, ensuring they remain effective against new attack vectors. Additionally, addressing the prevalence of false positives and negatives in detection is crucial. Future research should focus on developing sophisticated analytics techniques that enhance decision-making capabilities for cyber security professionals, reinforcing their invaluable expertise rather than overshadowing it.

Mobility and Infrastructure Evaluation

"Scenario Co-Design for Systemic Evaluation of Connected and Automated Mobility Setups" presents a compelling framework for enhancing the development and implementation of Cooperative, Connected, and Automated Mobility (CAM) services through participatory workshops that engage diverse stakeholders, including AV manufacturers, transport operators, infrastructure managers, and local authorities. This structured approach aligns technological innovations with community contexts, addressing critical gaps in existing methodologies. The research's impact is substantial, as it informs policy and strategic planning for urban mobility by promoting an inclusive decision-making process that reflects real-world challenges and opportunities. The practical applicability of the proposed framework is evident in its role as a tool for facilitating dialogue among stakeholders, creating a common knowledge base through detailed technical system descriptions, and addressing data-related challenges, such as trade secrets and data quality. Future directions for exploration include investigating the scalability of the workshop approach across different geographical contexts and refining mechanisms for data sharing to enhance evaluation reliability. Integrating advanced analytical methods, such as machine learning and real-time data processing, could further enhance the framework's applicability, enabling more dynamic mobility solutions. Overall, the article sets a foundation for future research and practice that aligns technological advancements with societal needs, contributing significantly to the field of smart mobility and urban planning." presents a compelling framework for enhancing the development and implementation of Cooperative, Connected, and Automated Mobility (CAM) services through participatory workshops that engage diverse stakeholders, including AV manufacturers, transport operators, infrastructure managers, and local authorities. This structured approach aligns technological innovations with community contexts, addressing critical gaps in existing methodologies. The research's impact is substantial, as it informs policy and strategic planning for urban mobility by promoting an inclusive decision-making process that reflects real-world challenges and opportunities. The practical applicability of the proposed framework is evident in its role as a tool for facilitating dialogue among stakeholders, creating a common knowledge base through detailed technical system descriptions, and addressing data-related challenges, such as trade secrets and data quality. Future directions for exploration include investigating the scalability of the workshop approach across different geographical contexts and refining mechanisms for data sharing to enhance evaluation reliability. Integrating advanced analytical methods, such as machine learning and real-time data processing, could further enhance the framework's applicability, enabling more dynamic mobility solutions. Overall, the article sets a foundation for future research and practice that aligns technological advancements with societal needs, contributing significantly to the field of smart mobility and urban planning.

Resilience in Water Systems

The article "Enhancing Resilience in IoT Water Systems Using Data- Intelligence and Decentralisation" represents a significant advancement at the intersection of cyber security and critical

infrastructure management, effectively addressing the growing concerns surrounding the security of water distribution networks — an area often overlooked in cyber security discussions. By integrating blockchain technology with advanced data validation mechanisms, the research enhances the resilience of water systems against malicious attacks and lays a foundational framework for incorporating intelligent decision-making processes, ultimately ensuring public health and safety. The practical usefulness of this research is manifold; the model specifications developed can be readily implemented across various platforms, including major cloud providers and blockchain networks, allowing for seamless integration into existing infrastructures. Additionally, optimizing assessment times and utilizing parallel processing techniques will significantly enhance operational efficiency, enabling water utilities to respond more effectively to real-time conditions and improving service delivery. Looking ahead, future research can explore optimizing node selection for blockchain validators, advanced detection techniques like federated learning, and expanding the range of attack scenarios examined within the EPANETCPA framework to mitigate risks from evolving attack vectors. Finally, enhancing the graphical user interface (GUI) is critical for ensuring the system remains user-friendly and accessible to non-technical stakeholders, facilitating better monitoring and control and promoting wider adoption of these technologies within the water sector.

Blockchain as a Service

Lastly, the article "Blockchain-as-a-Service: Architecture, Opportunities and Challenges" explores the critical aspects of Blockchain-as-a-Service (BaaS), emphasizing how cloud computing can enhance the scalability, efficiency, and accessibility of blockchain technology. Given the growing adoption of blockchain across various sectors, it highlights the necessity for a comprehensive understanding of its architecture and the associated opportunities and challenges when integrating blockchain with cloud services. By proposing solutions such as workload monitoring, minimizing interactions with blockchain nodes, and utilizing predictive resource allocation through neural networks, the research offers practical insights into optimizing the operational efficiency of blockchain applications. The emphasis on real-time communication, quality of service (QoS), and security implications related to the integration of IoT with BaaS contributes to the broader discourse on resilient and efficient blockchain solutions, particularly concerning scalability challenges in public versus private blockchains amid rising demands for high transaction volumes. Furthermore, the research provides actionable recommendations for developers and organizations implementing BaaS solutions, such as leveraging high-performance databases and caching layers to enhance performance and reduce API-related costs. The exploration of neural networks for predicting application workloads enables dynamic resource allocation, allowing organizations to adapt their cloud resources in real time based on anticipated demands, leading to improved user experiences and operational efficiencies. The article also identifies several future research directions, including dynamic resource allocation methods that leverage advanced AI and machine learning techniques, security enhancements addressing IoT integration challenges, a blockchain-based monitoring system for QoS to ensure compliance with service level agreements, and the effectiveness of layer 2 scaling solutions like sidechains, optimistic rollups, and zk-rollups to enhance public blockchain scalability. Additionally, investigating dynamic sharding techniques for private blockchains could provide significant insights into optimizing resource utilization and performance under varying workloads.

Through the above-elaborated articles, this SI aims to foster further exploration and collaboration among researchers and practitioners, ultimately shaping the future of IoT security and provisioning in these vital sectors. We hope that the contributions within this issue will inspire continued innovation and improvement in safeguarding our cyber-enabled critical infrastructure.

Biographies

Elias Bou-Harb [SM] (ebouharb@lsu.edu) received his postdoctoral training at Carnegie Mellon University and his Ph.D. degree in computer science from Concordia University, Montreal, Canada. He is currently an associate professor with the department of computer science at Louisiana State University, specializing in cyber security and data science as applicable to national security challenges. Previously, he acted as the director of the cyber center for security and analytics at the University of Texas at San Antonio, where he led and organized university-wide cyber security research, development, and training initiatives. Dr. Bou-Harb has authored more than 150 refereed publications in leading venues and has acquired significant state and federal cyber security research grants. His research and development activities focus on operational cyber security, cyber forensics, critical infrastructure security, empirical data analytics, digital investigations, network security, and network management. He is the recipient of seven best research paper awards, serves on various North American and International industry and university boards.

Edgar Weippl (edgar.weippl@univie.ac.at) is research director of SBA Research and full professor at the University of Vienna. Edgar's research focuses on fundamental and applied research on blockchain and distributed ledger technologies and security of production systems engineering. He (CISSP, CISA, CISM, CRISC, CSSLP, CMC) is member of the editorial board of Computers & Security (COSE) and associate editor of IEE Transactions on Information Forensics and Security (IEEE TIFS). He is Austria's representative at IFIP TC 11: Security and Privacy Protection in Information Processing Systems. He is steering committee chair person and involved in the organization of the ARES conference. He was General Chair of SACMAT 2015, PC Chair of Esorics 2015, General Chair of ACM CCS 2016, PC Chair of ACM SACMAT 2017, General Chair Euro S&P 2021 and 2024.

Chadi Assi [F] (assi@concordia.ca) is a Professor with the Concordia Institute for Information Systems Engineering at Concordia University, Montreal, Canada, where he currently holds a Tier I University Research Chair. He received his MASc and Ph.D. degree from the Graduate Center, City University of New York. Before joining Concordia University in 2003, he was a visiting scientist (2002–2003) at Nokia Research Center, Boston, working on quality-of-service in optical access networks. He received the prestigious Mina Rees Dissertation Award from the City University of New York in August 2002 for his research on wavelength-division-multiplexing optical networks and lightpath provisioning. He served on the Editorial Board of several IEEE journals and is currently serving as an Associate Editor for the *IEEE Transactions on Vehicular Technology*, *IEEE Transactions on Mobile Computing*, and *IEEE Transactions on Network and Service Management*. His current research interests are in the general areas of Networks (wired and cellular), network design and modeling, network optimization, and cyber security.

Jiangshan Yu (j.yu.research@gmail.com) is an Associate Professor at the University of Sydney. His research focuses on resilient decentralised systems, including blockchain technology. He has published in leading venues in the field and his research led to numerous competitive awards, including ARC DECRA (2021), IBM Academic Award (2020), and several best/distinguished paper awards, among others. He regularly serves in the Program Committee of leading conferences in security, system, and database, such as ACM CCS, USENIX ATC, IEEE/IFIP DSN, and VLDB. He is an elected Australian Country Member Representative to the International Federation for Information Processing (IFIP) Technical Committee 10 on Computer Systems Technology, and an elected member of the IFIP 10.4 Working Group on Dependable Computing and Fault Tolerance. He also serves as an Associate Editor for ACM Distributed Ledger Technologies (ACM DLT) and is on the Scientific Advisory Board for the Austrian Blockchain Centre (Austria). Jiangshan has delivered keynotes and invited presentations at prestigious conferences, universities, and industry events worldwide. His work has also garnered significant attention in news media globally.

Martin Husák (husakm@ics.muni.cz) is a researcher at the Institute of Computer Science at Masaryk University, a member of the university's security team (CSIRT-MU), and a contributor to The Honeynet Project. His Ph.D. thesis addressed the problem of early detection and prediction of network attacks using information sharing. His research interests are related to cyber situational awareness and threat intelligence with a special focus on the effective sharing of data from honeypots and network-monitoring.

Katherine A. Flanigan (kflaniga@andrew.cmu.edu) is an assistant professor of Civil & Environmental Engineering at Carnegie Mellon University. She also holds a courtesy appointment in Electrical & Computer Engineering. She has 7 years of experience researching structural health monitoring, cuber-physical systems, embedded systems, and communication technology in railroad, highway bridge, building, water system, and urban settings.

Join.
Connect.
Collaborate.

Make Our Community Your Community!

Join a global network of 30,000+ engineers, practitioners and academics working together to advance communications technology for the betterment of humanity.

As a ComSoc member, you will receive exclusive benefits to help you achieve your professional goals and stand out from your peers.

Benefits of a ComSoc Membership:

- Networking with Communications Technology Professionals Around the World
- Generous Conference Discounts
- Free Subscriptions to High-quality Technology Publications
- Top-notch Training and Continuing Education Resources
- Exciting Leadership and Volunteer Opportunities

IEEE ComSoc

www.comsoc.org/membership

The Role of Rule Mining in Generating Synthetic Cyber-Physical System Attack Samples

Merwa Mehmood, Zubair Baig, Naeem Syed, and Sherali Zeadally

Abstract

In recent years, Cyber-Physical Systems (CPSs) have increasingly been exposed to potential exploitation by the sophisticated adversary due to their vulnerabilities. The ever-evolving threat landscape for CPSs can impact their control logic, leading to system and process disruptions. Several Machine Learning (ML) based Intrusion Detection Systems (IDS) have been proposed to detect cyber threats in CPS. However, the issue of class imbalance in CPS datasets must be addressed to develop robust and effective security controls to mitigate cyber threats to CPS. We propose a novel method to generate synthetic ICS data by customising data generation methods specifically tailored for transactional datasets. The proposed scheme merges the process of mining frequent itemsets with a generative modeling method. A collection of items that frequently appear together is referred to as an itemset. We verify the validity of generated synthetic samples by comparing them with the original data samples. Furthermore, we apply three machine learning classifiers to evaluate the quality of the generated synthetic datasets with the aim to address the issue of class imbalance. The generated synthetic datasets to address the issue of class imbalance. The synthetic datasets generated contribute to the development of robust security controls which can detect evolving threats faced by CPSs.

Introduction

Industrial Control Systems (ICS) are a subset of CPS that comprise a diverse set of interconnected devices and processes operating in a coordinated way. The security of ICS is of paramount importance as they are often the target of numerous cyber-attacks aimed at penetrating and sabotaging operations and system controls. Stuxnet was one such attack on ICS that manipulated the control logic of industry field devices by altering the speed of operation of centrifuge Programmable Logic Controllers (PLCs) which compromise the integrity and confidentiality of the system [1]. Another common type of attack targeting the ICS is the False Data Injection Attack (FDIA) which manipulates the communication message between sensors/actuators and control units thereby disrupting the control logics of system.

One of the security measures used in an ICS is the use of model-based Intrusion Detection System (IDS) that detects threats and proactively issues warnings against suspicious activities or known attack signatures/patterns [2]. Signature-based attack detection schemes depend on known attack patterns and are generally ineffective in detecting attacks patterns that do not match known signatures. In contrast, several Machine Learning (ML)-based techniques have been proposed to learn the normal operations of the system and mark any deviations from it as anomalous. Due to the large amount of data generated in the ICS, applying ML techniques to learn the anomalous behaviour in communication messages is highly suitable. Typical ML based attack detection systems are trained to learn normal and malicious behavior/patterns acquired from ICS communication data. However, many ML-based IDS' have limitations in detecting variant attack signatures as training samples for attacks are scarce. The use of simulated attack datasets do not capture the complex interrelationships generally found in CPSs. Hence to build a CPS which is resilient to diverse cyber attacks, it is essential to address the issue of attack versus normal class imbalance.

Related Work

A major challenge in an ICS is the limited availability of real-world datasets for training and testing, which limits their suitability for real-world situations. There is a scarcity of literature pertaining to the Operational Technology (OT) aspect of ICS networks, creating a knowledge gap between the vulnerabilities and corresponding attack strategies. Over the past few years, the problem of data imbalance has been addressed by two methods, namely,

1. Sampling methods
2. Generation of artificial data

Random oversampling [3] duplicates the minority samples, leading to the repetition of information/samples in the training dataset. Random under-sampling reduces the majority class, resulting in biased outcomes. One of the widely used methods to generate artificial data is the SMOTE technique, i.e., synthetic minority oversampling [4], which involves interpolation between minority samples to generate new data samples. It per-

Merwa Mehmood, Zubair Baig, Naeem Syed are with Deakin University, Australia; Sherali Zeadally is with the University of Kentucky, USA and the University of Johannesburg, South Africa.

Digital Object Identifier: 10.1109/IOTM.001.2300272

forms well with low-dimensional data, but the high-dimensional nature of the ICS dataset affects the efficiency and performance of the model. The class imbalance issue could affect the performance of anomaly detectors by fostering a biased decision against the majority class.

A second type of solution proposed to deal with class imbalance is the generation of artificial or synthetic minority class. Generative Adversarial Networks (GANs) have been widely adopted for generating synthetic data samples. Especially in scenarios where data consists of images [5], videos [6] and text-to-image synthesis [7]. The successful generation of these types of synthetic datasets has demonstrated their suitability for generating realistic synthetic samples because they avoid sample repetition and perform well with high-dimensional data. However, GAN-based systems are suitable for generating realistic synthetic samples that are continuous in nature [8]. ICS datasets contain features that are both continuous and discrete, which can include sensor values as well as actuator states posing learning issues with existing generative techniques. To train GANs on large ICS datasets that consist of two types of data (continuous and discrete) remains a significant challenge, which reduces the efficient generation of synthetic samples for ICS scenarios. A previous work [9] describes how discrete features are managed, but it is hard to adapt to an ICS dataset because it is a combination of discrete and continuous data types.

To address this issue, in this work we focus on the rule-based learning technique specifically designed for transactional datasets that contain categorical features. Specially we focus on developing a detection framework that can detect FDIA in a water treatment system which combines an unsupervised machine learning method with a generative modeling approach. In our work we adopt the Association Rule Mining (ARM) algorithm (unsupervised machine learning) which uncovers hidden relationships in the dataset with the itemset generation algorithm (IGA) which generates synthetic FDIA samples. Techniques applied for generating synthetic samples often aim to produce samples that closely resemble the original datasets.

Association Rule Mining (ARM) is a technique for extracting relationships between feature categories that exist in large datasets [10]. ARM can identify patterns or relationships regarding how different items co-occur within a dataset. A collection of items that frequently appear together is referred to as an *itemset*. ARM generates frequent itemsets and association rules to present relationships between these itemsets. For instance, a 1-itemset represents the frequency of occurrence of a single item/category on its own, and a 2-itemset represents the frequency of co-occurrence between two items. If a low value of sensor A causes an actuator B to be in an OFF state, then the ARM model will identify a strong association between the "Sensor A: LOW" and "Actuator B: OFF" categories.

ARM uses two metrics, namely, confidence and support, to measure the strength of these rules. Confidence measures how often the rule is true. It is the probability that indicates how frequently Actuator B remains in OFF state when Sensor A has a LOW value. (e.g., Sensor: LOW). The output of the ARM process is both the frequent itemsets and the generated association rules. We filter out only the rules with support and confidence values above the user-defined thresholds as strong rules. In our scenario, we aim to use these frequent itemsets and association rules to create synthetic data because these rules highlight the strong relationships of various features and their observations. To generate synthetic samples that maintain the relationships found in the original dataset, the outputs of ARM are fed to a Itemset-based Generative Model (IGM). which filters the frequent itemsets based on a probability condition to determine which itemset to be selected for analysis.

Contributions of This Work

We summarize the main contributions of this work as follows:

- Discovering the hidden associations among features to understand the complex industrial processes and identify features that influence system behavior.
- Generation of synthetic samples using itemset-based generative model to address the class imbalance issue and ensure acceptable variance from the original dataset.
- We were able to establish relationship between 11 features with 54621 attack samples using the association rule mining technique. Synthetic attack samples enhanced ML model performance with a classification accuracy of 0.91.

Proposed Methodology

Figure 1 presents the proposed methodology shown. In this section, we proposed a method for the generation of synthetic false data injection attack (FDIA) samples. To generate synthetic samples and to seek a deeper understanding of attack patterns we integrate the rule mining technique (ARM) with the generative modelling technique (IGM).). This work is divided into two section one is Association rule mining engine that generates the frequent-itemsets and second one is Itemset generator that generates synthetic transactional dataset using the frequent itemsets as input from the association rule mining engine. The earliest work [11] related to integrating the frequent itemset with learning generative models explained the connections between mining frequent itemset with learning models. To generate synthetic transactional dataset using a generative model, we describe the relationship between frequent-itemsets and generative model by generating a transactional dataset of synthetic attack samples.

To generate a synthetic sample, step 4 of Algorithm 1 does discretisation of frequent itemsets. The IGM algorithm generates the synthetic transactional dataset, which is combination of pattern space and noise space. We obtain patterns or data samples randomly by selecting the frequent itemset values of operational sensor measurements that are different from the pattern space. This introduces a variability in the generated synthetic transactional dataset. For instance, the pattern space $T(X)$ = {Sensor A: HIGH, Actuator A: LOW}, this discretised sensor has range of values. In order to generate a noise sample $T(X')$, random sampling of sensor values in HIGH range generates a variability in the final transactional dataset.

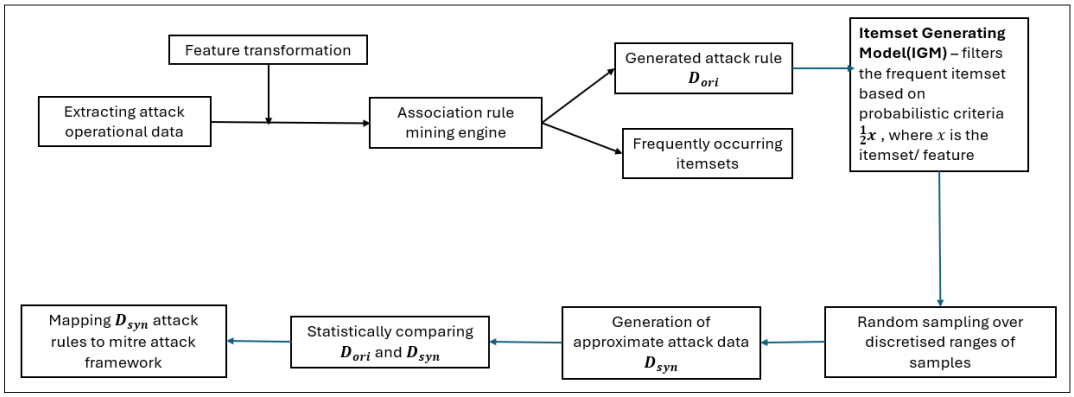


FIGURE 1. Schematic diagram of the proposed methodology.

Evaluation

Secure Water Treatment Plant System

We evaluated our proposed methodology for the Secure Water Treatment Plant system (SWaT). SWaT is a physical testbed of an ICS deployed at the Singapore University of Technology and Design [12]. This deployment collected attack and normal operational data over a period of 11 days of running in emulated mode. Normal data was collected without any disruption over 7 days and the next 4 days comprised the collection of 36 variants of false data injection attacks. There are six stages of water treatment for this plant and a total of 449919 samples representing 51 features with one target column representing the sample label as attack or normal, were generated. A timestamp column represents the duration of each attack sample. There are 54621 attack samples, and the remainder are normal samples. SWaT Stage 2 is a water pre-treatment phase that handles the chemical properties of water and controls the addition of chemicals to it.

Association Rule Mining Engine

We extract the 11 attributes of Stage 2 of the SWaT system which consists of sensor measurements and actuator states, namely, FIT201, MV201, AIT201, P201, P202, AIT202, P203, P204, AIT203, P205, P206. The three backup pumps P202, P204 and P206 were excluded, as there was no change in their state (values) throughout the dataset. Backup pumps are used to operate in conditions when the primary pumps (P201, P203, P205) fail but during the collection of these attack and normal samples, these pumps remain in a constant state. We have selected FIT201, MV201, AIT201, AIT202, AIT203, P201, P203, and P205 as the core features for our analysis. These features represent anomalies in terms of variation to legitimate data samples and are highly correlated to attack samples.

The dataset we selected for analysis is a combination of binary (e.g., pumps), real-valued (e.g., sensors) and ternary (e.g., motorised valve) features. However, Association Rule Mining (ARM) accepts only binary-valued attributes. The presence or absence of any feature in the transaction is represented as 0 or 1. For this reason, extracted features are discretised into specific design threshold values. The threshold values (conductivity analyser is a sensor measurement which measures the level of a chemical content in water, it has two levels: one is high level C_h = 260 and the low level C_l = 250 above and below these values indicating it as an attack) for real-valued features are converted into High, Medium, and Low categories. For the actuator state, we only considered low and high categories representing the closed (1) and open (2) states of the actuator respectively. The feature MV201 has an additional transition state, and it is binarized into 3 categories. These binarized values are then converted into encoded data and input to the rule mining engine. The attack model of the ICS dataset is a data tuple that is composed of sensor and actuator pairs representing an attack. This system's known attack models are similar to our proposed methodology. For instance, if the actuator is closed, the concentration of a chemical increases making the water acidic resulting in an attack. Here, we have opted for Association Rule Mining (ARM) as the technique to help find hidden association among features. For this, we integrated the rule mining technique with the generative modeling technique to generate synthetic samples.

The SWaT dataset consists of both attack and normal operational data. We extracted the attack data and performed attribute transformation on each feature. The discretised features are then used as an input to the rule mining engine. The rule mining engine generates the frequently occurring itemset, which is calculated using the FP growth algorithm(an efficient algorithm used to find frequent itemsets, it follows a tree like data structure which captures the frequency and associations of itemsets), with a minimum support threshold value of 0.0039 according to the probabilistic criteria of $1/2^x$, where x is the number of itemsets or features. We obtain approximately 1251 frequent itemsets from this minimum support value, wherein, 9 frequent itemsets were size 8-itemsets. After extracting these 9 itemsets, 1242 frequent itemsets were left. The selection of a minimum threshold is an important key factor in evaluating the frequent-itemset efficacy. A low threshold will generate many itemsets, and a high threshold value will eliminate some important frequent itemset patterns. To overcome this issue, we generated frequent itemsets by using the generative learning model but extracted itemsets of size 8 so that itemsets selected match with the input sample sizes. These frequently occurring itemset/features are patterns of interest and each frequent itemset is

considered as a synthetic dataset on its own.

Table 1 shows three variants of attack operational datasets. Additional synthetic datasets can be generated by following the frequent itemset generation technique. To generate synthetic samples, we used the frequent-itemset with length equal to eight according to the probability criteria mentioned in the step 5 of Algorithm 1. The reason for selecting an eight-itemset is because it resembles the input order/data sample size of the original dataset.

For instance, S_1 generates 8-itemset synthetic features that resembles the original dataset. We generate the synthetic dataset with continuous sensor measurements and numerical actuators states (of operational data). We have eight features in the synthetic dataset wherein three are sensors measurements that are real-valued features, and three actuator states as binary values. The output of IGM is a synthetic transactional dataset which cannot be directly combined with original data which includes both discrete and continuous features. Hence, sensor-based features in the discrete transactional dataset $T(X')$, need to be converted into the continuous format. To generate $T(X')$, we performed random sampling on the sensor measurements. In our first synthetic sample, the attack was generated when the actuator states were manipulated from an open (HIGH) (2) state to a closed (LOW) (1) state. We then performed random sampling of the sensor measurements that resulted in novel attack variants.

In transactional data S_1 ("AIT202(t)_High, " "P205(t)_Low, " "P201(t)_Low, " "P203(t)_Low, " "AIT203(t)_Medium, " "MV201(t)_Low, " "AIT201(t)_Low, " "FIT201(t)_Medium"), we added randomness according to the discretised ranges of the original dataset. The actuators are kept in the same state, which is closed, and the random sampling is performed on the high range category of AIT202 and is repeated for other binarized features. The reason for keeping the actuators in a closed state is that data preprocessing [13] yielded a high correlation among the three actuators. This indicates that whenever there was an attack, the actuators were manipulated from an open (2) to a closed state(1), disrupting the values of the chemical analysers(AIT201, AIT202, AIT202).This shows that the synthetic samples S1, S2 and S3 shows similar behavior as the original attack samples. We generated 500 synthetic samples for each sensor feature and compared it with the original sample's sensor measurements. To visualize the synthetic and original samples, Fig. 2 shows a comparison between the original samples and the synthetic attack samples. This dataset is an approximate dataset because it is generated from a discretised range of the original features. This binarization results in the loss of information. To further validate these synthetic samples, we considered the generated rules from the original and the synthetic samples and compared the similarity between them. Figure 2 shows the distribution plot of generated synthetic and original samples. The x-axis represents the value ranges of the sensor values, and the y-axis represents the frequency of occurrence of samples in the dataset. For instance, AIT203 distribution plot which is an ORP analyser the highest range for this chemical is above 340, during an attack the actuator associated with this analyser is in closed state that affected the concentration of chemical as the plot.

D_{syn}	Generated frequent itemset
S_1	'AIT202(t)_High', 'P205(t)_Low', 'P201(t)_Low', 'P203(t)_Low', 'AIT203(t)_Medium', 'MV201(t)_Low', 'AIT201(t)_Low', 'FIT201(t)_Medium'
S_2	'AIT202(t)_High', 'P205(t)_Low', 'P201(t)_Low', 'P203(t)_Low', 'MV201(t)_Low', 'AIT203(t)_Low', 'AIT201(t)_Low', 'FIT201(t)_Medium'
S_3	'AIT201(t)_Low', 'MV201(t)_High', 'P203(t)_Low', 'FIT201(t)_High', 'P205(t)_Low', 'P201(t)_Low', 'AIT202(t)_Low', 'AIT203(t)_Low'

TABLE 1. Generated frequent itemset.

1. Collect attack operational data.
2. Attribute transformation according to design thresholds.
3. Generate attack rules by association rule mining engine D_{ori}.
4. Generate frequent itemset (χ) using discretised features.
5. Frequent-itemset model generates synthetic dataset based on probabilistic criteria of $(1/2\chi)$.
6. Itemset-generator generates synthetic transactional dataset $D_{Syn} . = T(X) \cup T(X)$

ALGORITHM 1. ICS data synthesizer.

Validation of Synthetic Samples

We validate our generated synthetic samples using two evaluation criteria. The first criterion compares the interdependency between the features and the observation of common attack patterns/rules in both datasets. The second criterion involves performing a machine learning based prediction test that evaluates how well the synthetic samples generated by the generative model align with the original dataset.

Attack Rule Generator

We used a rule-based machine learning technique on real and synthetic datasets to discover the strong association among features. Our aim is to identify whether a specific association in real datasets is also present in the original dataset. To achieve this, we applied the rule mining technique on our original dataset and the synthetic dataset. For simplicity we only focussed on rules having a length of 2, e.g., antecedent part having 2 features only. The output of the generated association rule looks like A -> B, wherein, one rule is (P201_LOW -> AIT202_HIGH) having P201_LOW as an antecedent and AIT202_HIGH as the subsequent part of the rule. If P201_LOW implies actuator P201 is turned off, then the value of AIT202 will always be high for a data injection attack. We extracted approximately 90–110 association rules from each of these eight features for generating the synthetic data. From the original dataset, we extracted around 80–100 rules.

One of the methods to evaluate the generated synthetic datasets is by comparing the predictive and the recall abilities of the models when trained on both real and synthetic datasets. Precision is a positive predictive value, which shows the correctly classified positive instances and Recall identifies the true positive classes identified by the model. A good quality synthetic dataset is expected to give similar predictive and recall abilities across all ML models (We have used only precision and recall to evaluate the strength of the generated synthetic samples). Our proposed method includes an evaluation step to statistically verify the generated samples. In our proposed model we evaluate the synthetic rules generated by the rule engine and only selected samples with support values of 5% and 10% along with a confidence value

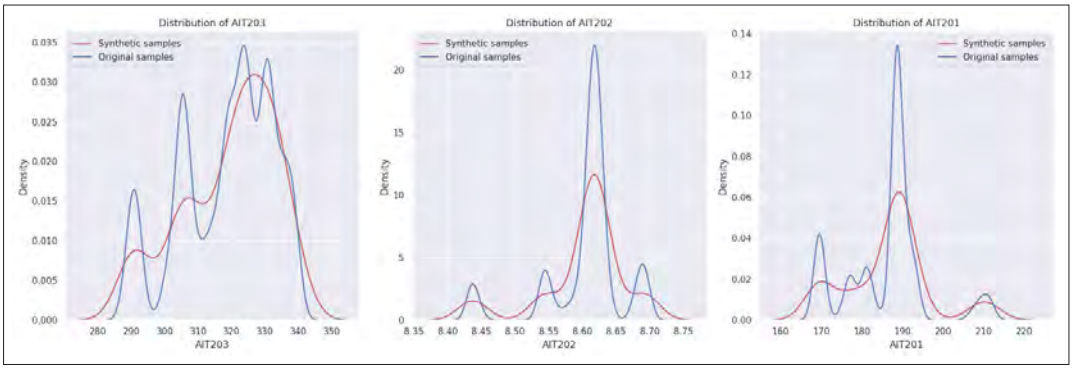


FIGURE 2. Distribution plot of three chemical analysers (AIT201, AIT202, AIT203).

Dataset	Generative model	Number of rules extracted from real data	Number of rules extracted from synthetic data	Number of matched rules in both datasets	Number of matched frequent item sets in both datasets	Precision	Recall
SWaT 2015	IGM	88	110	68	35	0.61	0.77

TABLE 2. Association rule mining generator results.

Dataset	Logistic regression	Support Vector Machine	Random forest
D_{syn}	0.923	0.948	0.99
D_{ori}	0.93	0.943	0.98

TABLE 3. Prediction performance results.

of 40–50%. There were a high number of rules that were similar in both real and original samples. Table 2 shows a sample evaluation result. The high recall value demonstrates the similarity between the synthetic sample and the original dataset. The precision shows relevance around 60 percent in the dataset, that validate the generated synthetic samples. It also shows that the manipulated actuator state from open to closed state is constant for these evaluation results.

Time series analysis of stage-2 features helps us to understand anomalous behavior and its impact on the sensors and actuators in the system over a time period and explains the disturbances in the normal periodic pattern. A detailed time series analysis is published in our previous work [14], which explains the interdependences between the chemical analysers (AIT201, AIT202, AIT203) and their actuators (P201, P203, P205). We note from generated synthetic samples S1, S2, S3 that the closed state of actuators affected the concentration of chemical analyser values. For instance, in S1 ("AIT202(t)_High")(pH analyser) means AIT202 > 8 value as Figure 2 shows, "P203(t)_Low" is in closed state increasing the concentration of chemical and pH of water making unfit for consumption. The MV201 is interdependent with FIT201 flow transmitter values associated with the flow of the water into these three chemical tanks. So, we focused on these 6 features for generating the synthetic samples. It is also worth noting the behavior of MV201 and FIT201 across three synthetic samples; when the MV201 is closed or low the FIT201 is in medium range that is between the low and high state that may be stabilising the flow when the valve is close and similarly in S3 the MV201 is open the FIT201 is also high.

Predictive Modelling on Synthetic Data

To solve the issue of class imbalance in an ICS dataset and to evaluate how well the synthetic samples are generated through the predictive modeling technique, we addressed the robustness of synthetic transactional dataset when compared with the original data samples. We generated approximately 500 synthetic attack samples using the frequent itemset generating model and augmented it with the original attack samples. By combining real and generated attack samples we obtained a total of around 55000 attack samples for our system. To validate the efficiency of both datasets, we evaluated the predictive performance of three machine learning classifiers. Table 3 shows that for Logistic regression, Support Vector Machine (SVM) and Random Forest, we obtained very similar classification accuracies. This shows that the generated attack variants appear to be realistic and can be incorporated into the original dataset to resolve the issues of class imbalance.

MITRE ATTACK Framework

In this section we present a real-world context for artificially generated synthetic attack samples. The generated synthetic sample causes potential vulnerabilities in the ICS which, can be exploited by an adversary. The attacker can use the synthetic sample generated to carry out a real-world attack. We leverage MITRE ATT&CK tactics that can be utilised by the attacker to compromise the ICS system [15].

We evaluated one of the attack variants generated by our proposed methodology for the water treatment plant (SWaT) system. An attack scenario to this water treatment stage involves increasing the concentration of a chemical by manipulating the actuator state. There are three chemical analysers that check the concentration of a specific chemical in the water. These analysers are controlled by their corresponding actuators. Manipulating any of the actuators would result in an increase in the chemical concentration resulting in an attack. The ATT&CK framework has an ICS matrix that addresses the IT-OT relationship, by launching an attack against the IT part (e.g., accessibility to Internet-connected device) and its impact on the OT part of

Tactic	Technique	Description
TA0108 — initial access	T0883 — Internet accessible device	Accessing the targeted assets(workstation) to which the PLC is connected
TA0104 — execution	T0821 — modify controller task	Manipulating the data of field devices (PLC) using Adversary-in-the-middle (AiTM) attacks (T0830) modifying the parameter with false actuator state
TA0103 - evasion	T0856 —spoofing reporting message	Adversaries manipulate the reporting message and impair the control process by blocking the reporting messages(T0804)

TABLE 4. Adversarial tactics and techniques.

the system by causing damage to the control processes. Table 4 presents the tactics and techniques adopted by the adversaries to execute an attack.

We have addressed three tactics and techniques that can be adopted by an adversary of an ICS. The attack starts with the initial access of the network, gaining insights about the field devices. The second tactic is the execution step that aims to modify tasks of the controller acting as an adversary in the middle (Man in the Middle) to manipulate the device logic. The third tactic is evasion by generating spoofing messages and blocking messages generated by the field devices for delivery to the PLC. This includes spoofing of a reporting message (T0856) that affects the control process. The reporting message contains information about the input and output values. If an adversary has control over a communication message, then the adversary can make changes to the control logic without being detected by the system. In our scenario, the generated synthetic sample ("P205(t)_Low," "P201(t)_Low," "AIT203(t)_Low," "AIT201(t)_Low," "P203(t)_Low," "AIT202(t)_High") is an attack variant that we align with the MITRE ATT&CK framework. The adversary gets access to a PLC that controls the logic of field devices. Each stage has two PLCs (a primary and a secondary. By gaining access to the appropriate PLC, an adversary can execute the Adversary in the Middle (AiMT) attack by spoofing the actuator state resulting in an increase in the chemical concentration in the water and subsequently causing human harm when the modified water is consumed.

Conclusion

At present, there is limited data available for representing attacks against an ICS. We have addressed the issue of class imbalance through the proposal of a novel rule mining-based technique, for attack modeling and for the synthesis of attack data samples. Analysis of the results obtained using the proposed data samples, when trained against three popular classifiers, demonstrates the efficacy of our proposed technique for data sample synthesis.

Acknowledgment

We thank the anonymous reviewers for their valuable comments which helped us improve the content, quality, and presentation of this article. Sherali Zeadally was partially supported by a Distinguished Visiting Professorship from the University of Johannesburg, South Africa.

References

[1] S. Collins and S. McCombie, "Stuxnet: the Emergence of A New Cyber Weapon and Its Implications," *J. Policing, Intelligence and Counter Terrorism*, vol. 7, no. 1, 2012, pp. 80–91.

[2] M. Jin *et al.*, "Boundary Defense Against Cyber Threat for Power System State Estimation," *IEEE Trans. Information Forensics and Security*, vol. 16, 2020, pp. 1752–67.

[3] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for dealing with Imbalanced Classification in Educational Data mining," *Information*, vol. 14, no. 1, 2023, p. 54.

[4] D. Elreedy, A. F. Atiya, and F. Kamalov, "A Theoretical Distribution Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Imbalanced Learning," *Machine Learning*, 2023, pp. 1–21.

[5] T. Karras *et al.*, "Training Generative Adversarial Networks with Limited Data," *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12,104–14.

[6] C. Roberto de Souza *et al.*, "Procedural Generation of Videos to Train Deep Action Recognition Networks," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 4757–67.

[7] J. Agnese *et al.*, "A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 4, 2020, p. e1345

[8] S. K. Alabugin and A. N. Sokolov, "Applying of Generative Adversarial Networks for Anomaly Detection in Industrial Control systems," *2020 Global Smart Industry Conf. (GloSIC)*, Nov. 2020, pp. 199–203.

[9] R. V. Raj, V. Sangeetha, and P. P. Amritha, "GAN-Based Anomaly Intrusion Detection for Industrial Controller System," *World Conf. Information Systems for Business Management*, Sept. 2023, Singapore: Springer Nature Singapore, pp. 79–89.

[10] R. Agrawal, T. Imieli?ski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," *Proc. 1993 ACM SIGMOD Int'l. Conf. Management of Data*, 1993.

[11] S. Laxman *et al.*, "Connections Between Mining Frequent Itemsets and Learning Generative Models," *7th IEEE Int'l. Conf. Data Mining (ICDM 2007)*, Oct. 2007, pp. 571–76.

[12] Y. Ji *et al.*, "Adversarial Attacks and Mitigation for Anomaly Detectors of Cyber-Physical Systems," *Int'l. J. Critical Infrastructure Protection*, vol. 34, 2021, p. 100452.

[13] M. Mehmood, Z. Baig, and N. Syed, "Securing Industrial Control Systems (ICS) Through Attack Modelling and Rule-Based Learning," *2024 16th Int'l. Conf. COMmunication Systems & NETworkS (COMSNETS)*, Bengaluru, India, 2024, pp. 598–602. DOI: 10.1109/COMSNETS59351.2024.10426882

[14] M. Mehmood, Z. Baig, and N. Syed, " Time Series Analysis and Rule Mining for Detecting Industrial Control System Data Injection Attacks," (unpublished work) *2024 Symp. Emerging Topics in Computing and Communications (SETCAC'23)*, Bengaluru, India.

[15] M. D. Firoozjaei *et al.*, "An Evaluation Framework for Industrial Control System Cyber Incidents," *Int'l. J. Critical Infrastructure Protection*, 2022, vol. 36, p.100487.

Biographies

Merwa Mehmood (mehmoodme@deakin.edu.au) is a Ph.D. candidate in Cybersecurity with the Strategic Centre for Cyber Resilience and Trust (Deakin CYBER), Deakin University.

Zubair Baig (zubair.baig@deakin.edu.au) is an Associate Professor in Cybersecurity with the Strategic Centre for Cyber Resilience and Trust (Deakin CYBER), Deakin University.

Naeem Syed (naeem.syed@deakin.edu.au) is a Lecturer in Cybersecurity with the Strategic Centre for Cyber Resilience and Trust (Deakin CYBER), Deakin University.

Sherali Zeadally (szeadally@uky.edu) is a Professor in the College of Communication and Information at the University of Kentucky.

Toward Intelligent IoT Endpoint Detection and Response Using Digital Twins via Firmware Emulation

Shin-Ming Cheng, Yi-Ching Lui, Nien-Jen Tsai, and Bing-Kai Hong

Abstract

The properties of short time-to-market, heterogeneity, constrained resources, and unfriendly interfaces for IoT endpoint devices render system-based security mechanisms in traditional desktops, such as antivirus, inapplicable. Moreover, popular network-based security solutions, such as IDS, might not completely detect and mitigate the rising fileless IoT attacks. This article leverages recent innovation, firmware emulation, to enable a digital twin (DT) of a targeted actual IoT endpoint device and to realize an intelligent IoT endpoint detection and response (EDR) platform. Inbound traffic to the actual IoT endpoint device is mirrored to the DT in the platform, and the system-level monitoring module integrated into the softwarized DT provides deep IoT endpoint detection in ways that are not possible on physical IoT endpoint devices. Machine learning algorithms are proposed to identify malicious behavior from system calls and network packets collected from system-level and network-level monitors, and suspicious packets containing harmful commands are further determined. The EDR consequently updates the IDS rules so that traffic to the actual IoT endpoint device with the same malicious patterns is recognized and blocked, thereby achieving endpoint response. In the experiment, we enable emulation of IoT endpoint devices with ARM, MIPS, and X86 architectures and realize Mirai malware and remote code execution (RCE) attacks to validate the proposed EDR platform. With a 99.94% accuracy rate in attack determination, we believe that the proposed solution is feasible for the protection of IoT endpoint devices behind the edge. Such outcomes identify secure functionalities that DT using firmware emulation could offer in the IoT paradigm, thereby opening the door to innovative mechanisms to combat IoT attacks.

Introduction

With sensing, computing, and communication capabilities, the Internet of Things (IoT) bridges the physical world and cyberspace, providing various kinds of applications to humans. The IoT framework comprises endpoint sensors and actuators (known as IoT endpoint devices), infrastructure edge nodes for data relaying (known as edge nodes), and IoT application servers. The fact that any modification to IoT devices in the cyber world will affect users' safety and privacy makes IoT a valuable target for adversaries. Moreover, IoT applications are typically designed for specific purposes, and instead of security requirements, aspects such as cost, performance, or power are the main considerations during the design process. For example, the short time-to-market increases the possibility of hard-coded passwords, thereby making IoT endpoints vulnerable to malware infection and fileless attacks [1]. Consequently, the security issue in IoT has been an ever-increasing concern.

Without sufficient computing power and a convenient access interface, traditional well-developed *host-based* protection mechanisms (e.g., antivirus) cannot be directly shifted into the IoT paradigm. *Network-based* solutions located at edge nodes such as a firewall or intrusion detection system (IDS) are feasible, where flow and packets to/from IoT endpoint devices are monitored and malicious ones are identified and blocked [2]. Various kinds of features, such as traffic volume, IP address, and port number, flow semantics, or payload and data, are extracted for the estimation of malicious traffic or malfunctioning IoT devices [3]. The powerful machine learning (ML) algorithms are applied as inference techniques from measured features to detect unknown malicious traffic automatically [2].

Although ML-based network detectors remarkably improved the defense performance of IoT endpoint devices, the fundamental issue that only network traffic can be utilized for analysis still restricts the degree of protection. E-Spion [4] first leverages system-level information such as CPU utilization or system call during the execution of the target process to determine the abnormal behavior. However, E-Spion suggests that the physical hardware of an IoT device should be connected to the IDS, which introduces a hardware dependence issue and is not scalable. By virtually rehosting firmware into an emulated IoT system, the operations of firmware are virtualized and decoupled from the original IoT endpoint hardware. The barriers of constrained resources and inaccessibility in IoT endpoint devices are tackled accordingly [5]. The emulated IoT endpoint being enabled with powerful computation capability could operate exactly the same as the original IoT endpoint hardware. Acting as a virtual replica, the

The authors are with National Taiwan University of Science and Technology, Taiwan; Shin-Ming Cheng is also with the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. Nien-Jen Tsai is the corresponding author.

Digital Object Identifier: 10.1109/IOTM.001.2400070

emulated IoT endpoint could reflect the current status in high fidelity and is considered as a Digital Twin (DT) of the original IoT endpoint hardware [6].

By integrating system-level monitors in virtualized DT, the system behavior during operation such as system call or instruction executed can be well captured [7], which complements network-based solutions. Rather than directly enabling dynamic analysis and assessment in the emulated IoT system (e.g., fuzzing and concolic execution), this article achieves system-level detection in emulated IoT devices so that the original IoT endpoint is protected. In particular, the system-level monitors in the emulated IoT system apply an ML algorithm to identify the malicious action and determine the corresponding network traffic or packets. Then the edge could easily block suspicious traffic incurring the malicious activity so that IoT endpoint devices behind the edge are protected. We believe such IoT endpoint detection and response (EDR) is the first feasible solution without the support of IoT hardware and thus is much more scalable and efficient. We believe that introducing an emulated IoT system that acts as DT paves the way to designing practical defenses and future research that uses this as a foundation are expected to be developed as new security solutions for IoT endpoint devices.

The contributions of our study are as follows,

- **Digital Twin Framework.** Designed and implemented a digital twin framework that uses firmware emulation to create virtualized replicas of IoT devices for security analysis.
- **Enhanced System-Level Monitoring.** Advanced system-level monitoring capabilities with Strace, Mshell, and SystemTap leverage digital twins to detect and respond to potential security threats in IoT devices.
- **Firmware Vulnerabilities Identification.** Identified three previously known vulnerabilities in IoT firmware through testing and analysis using the digital twin framework.

The remainder of this article is organized as follows. We survey the existing research in network-based edge detectors and DT. The core technology applied to enable CDT, firmware rehosting, is extensively described. The novel intelligent IoT EDR based on the emulated IoT system is proposed and the experimental results are discussed. Finally, we conclude this work.

Background and Related Work

IoT Attacks

Recently, IoT botnets and malware received lots of attention due to the Mirai's source-code release and damage on global websites from its variants. Moreover, hackers exploit existing or unknown vulnerabilities on the victim devices to achieve fileless attacks without transporting a malicious binary [1]. The current IoT attacks consist of following stages [8]:

Stage 0: Scanning. By investigating reactions of crafted requests, the adversary could locate and identify vulnerable victims for the following attacks.

Stage 1: Exploitation. It is typically achieved by brute force password guessing or exploiting RCE vulnerabilities to gain access to the victim.

Stage 2: Downloading. The malicious binary is delivered from a loader to the victim.

Stage 3a: Execution. The victim is compromised via the execution of payload.

Stage 3b: Compromise. The adversary completely takes over the full control of the victim and could manipulate the victim persistently.

Stage 4: Communication. The victim communicates with the Command and Control (C2) server or the adversary and receives instructions from them.

Stage 5: Attack Action. The compromised victims acting as bots or relays launch stealth attacks such as distributed denial-of-service (DDoS) or lateral movement.

Before launching a malware attack, an adversary first collects publicly available IoT devices using a search engine such as shodan, which is often referred to as reconnaissance (see Step 0). The infection is typically achieved by brute force password guessing (see Step 1). In the downloading stage, the adversary typically leverages `wget`, `curl`, or `echo` commands to download malware to the victim (see Step 2), and then execute malware to infect the IoT device to form a botnet (see Step 3a). Unlike a fileless attack, an adversary will use the infected victims to connect to a C2 server (see Step 4), so that the hacker can control a large number of infected zombies via C2 server to launch DDoS attacks against specific services in a short period of time (see Step 5).

Regarding fileless attacks, the adversary deliberately hides their actions using known RCE vulnerabilities and leaves no files behind (see Step 1), thereby increasing the difficulty for later forensics [1]. Subsequently, the adversary can implant backdoors into the chosen victim devices to execute advanced persistent threats (see Step 3b). Such attacks are highly suitable for conducting subsequent attacks such as privilege escalation, data theft, information exposure, or network compromise.

IoT Network-Based Detector

With powerful computing capability, detectors located could monitor the communication traffic from/to the targeted IoT endpoints in real-time [2]. Such network-based IDS extracts features from the packet header, payload, or network flow and further recognizes specific activities from the measured features [3]. Traditionally, activities are compared with malicious ones in a signature database predefined by security experts. If there is a match, then the activity is determined as suspicious. For example, Heimdall [9] leverages whitelist and blacklist queries from VirusTotal as the basis for determination.

To detect unknown malicious traffic automatically, ML-based solutions with relatively high accuracy and a low false alarm rate have been proposed [2]. FlowGuard [10], for instance, inspects all packets passing through and identifies DDoS traffic, which is then blocked to protect IoT endpoint devices behind the edge. The Long Short-Term Memory (LSTM) ML technique is applied because temporally correlated DDoS traffic can be precisely captured. Passban IDS [2] learns the system's normal behavior during the training phase and then detects anomalies in incoming network traffic, particularly DDoS attacks. Additionally, to identify malicious traffic other than attack actions (i.e., stage 5), flow semantics are analyzed by investigating several consecutive interactions.

Traditionally, activities are compared with malicious ones in a signature database predefined by security experts. If there is a match, then the activity is determined as suspicious.

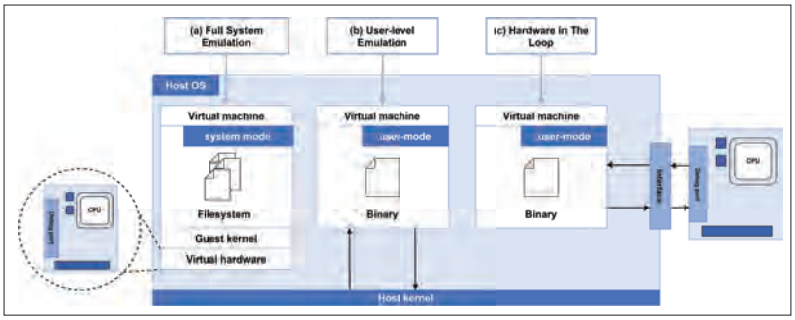


FIGURE 1. Firmware Emulation Techniques.

Digital Twin

Originated from intelligent manufacturing, a virtual twin that digitally projects a physical entity receives data from the physical counterpart and replicates its behavior [6]. With the aid of output from the virtual representation, real-time monitoring and controlling, fault diagnostics and early prediction, or dynamic optimization of the asset are enabled.

DT has become increasingly prevalent in the realm of IoT, serving as virtualized counterparts to sensors, actuators [11], and other IoT devices. They find application across various domains including healthcare [12], where they simulate digital patients, as well as in intelligent vehicles and IoT device management [13]. For instance, through the execution of representative functionalities and close integration with physical sensors, software replicas can closely mimic real-world behavior. Consequently, DTs serve as specialized logical entities tailored to specific IoT applications [11]. Leveraging in-body, on-body, and environmental sensors along with affordable devices, it becomes feasible to create digital representations of patients linked to targeted individuals. By harnessing data collected from these sensors, it becomes possible to discern the activities experienced by the individual, thereby facilitating improved care outside traditional healthcare settings and enabling the practice of "precision medicine" [12].

In order to handle the enormous amount of data measured from the physical object, various kinds of ML algorithms are proposed in the virtual twin [14]. In particular, a predictive model is responsible for predicting information using ML algorithms or neural networks so that further decisions are made and trade-offs are analyzed [6].

Firmware Rehosting for IoT Cybersecurity

By isolating execution from co-located physical hardware, emulation is now becoming a popular tool for software development, security analysis, and logic debugging [5]. The firmware of the targeted IoT endpoint device is extracted and rehosted within an emulation environment using three main approaches below.

User Program Emulation: As shown in Fig. 1b, the single binary of a particular service is executed as a process using QEMU user mode without emulating the entire firmware, kernel, and peripherals. A CPU emulator with a virtual stack and memory segment interprets and executes each instruction decoded from the binary. While process-level emulation proves efficient and apt for intricate security tests on a single binary like fuzzing, it does have drawbacks. Operations related to peripherals can lead to unexpected failures.

Hardware in The Loop (HITL): As illustrated in Fig. 1c, the CPU emulator in HITL can receive actual hardware responses when executing relevant instructions, as they are forwarded to the real IoT hardware via the debug port. However, hardware dependence inevitably reduces scalability and parallelism.

Full System Emulation: We can alternatively fully emulate the kernel, filesystem, and common peripherals using QEMU system mode, as depicted in Fig. 1a. The advantages of independence and high fidelity come with the cost of laborious and error-prone manual configurations, given the heterogeneous architectures and peripherals in IoT devices. Addressing the demand from a vast and rapidly increasing number of IoT devices, researchers have been focusing on automated emulation solutions.

DT Framework in Intelligent IoT EDR

Figure 2 depicts the proposed intelligent IoT EDR as a DT framework, aimed at monitoring, identifying, and thwarting malicious behavior at the system level. The DT is structured into data and model components. In Fig. 2a, the data component primarily comprises emulated IoT devices facilitated by rehosting technology, a topic thoroughly discussed later. To comprehend the behavior of these emulated devices, we've integrated a system-level monitoring module, detailed later. Network traffic bound for the actual IoT device is mirrored to the virtual DT's data component for thorough system and network level scrutiny. The data extracted from this analysis is then fed into the model component for behavioral analysis.

In Fig. 2b, the malicious behavior detector focuses on scrutinizing system calls for abnormal commands and labeling irregular log files. Additionally, the command extractor depicted in Fig. 2c identifies commands within the abnormal logs, translates them into IDS rules, and deploys these rules to the EDR. Subsequent traffic exhibiting similar attack patterns will be promptly identified and blocked based on these established rules.

Rehosted Firmware in DT

An emulated firmware is considered a DT of the actual IoT endpoint device since it operates exactly the same as the actual device. With sufficient computing power, we opt for full system emulation due to its high fidelity. To build the virtualized DT, we first extract the firmware of the actual IoT device and then rehost it using a well-known emulation tool, Firmadyne [15]. The firmware datasets we used are the same as those in the Firmadyne dataset package. We extracted the firmware using Binwalk, which involved unpacking compressed archives, extracting file systems, and isolating embedded files. This process allows for smooth acquisition of the filesystem and necessary modifications. Once we confirm the firmware architecture, we replace the original kernel to support essential system tools. Subsequently, the system is simulated using QEMU system-level emulation with appropriate configurations. Finally, the network interface undergoes intensive configuration, resulting in the successful establishment of the DT.

Ensuring compatibility is crucial when integrating system-level monitoring modules into emulated firmware. This involves testing and adjusting the kernel accordingly. Additionally, the diverse range of IoT

devices complicates automated firmware emulation. It's essential to thoroughly examine the various versions of libraries utilized in the system. Sometimes we need to compile the relevant tools with a suitable cross-compiler to ensure compatibility.

We present the first firmware emulation research within the context of DT, as compared to other studies. To assess fidelity in firmware rehosting, the study [5] examines various methods and their respective levels of fidelity. While our emulation achieves only module-level fidelity, we enhance the automation of our system, thereby improving scalability for further analysis. For instance, rather than prioritizing fidelity improvements, we might slightly modify the booting and kernel-related configuration to match the virtual environment. This is advantageous because some analysis tools can only run on specific versions of the kernel.

System-Level Monitoring

In contrast to network-based IDS solutions, which can only monitor network traffic, the proposed DT framework incorporates a system-level monitoring module. This enables the comprehensive capture of system-level behaviors outlined in Steps 1 and 3 in and earlier section. We specifically use the following three approaches for system-level monitoring: Strace, Mshell, and SystemTap. Strace and Mshell operate in user-space, while SystemTap operates in kernel-space. Detailed explanations are provided below.

Strace: As depicted in Fig. 3a, this is a typical debugging tool utilized for tracking system calls in progress. It operates at the user level and is implemented through the underlying `ptrace` probe. This tool allows us to observe the behavior of a specific process by hooking into the IoT device. All system calls performed by the process are recorded for subsequent analysis.

Middle Shell (Mshell): The shell serves as a user interface for interacting with the underlying system, enabling users to execute commands. Our modification involves transforming the default shell into an Mshell, capable of logging all commands intended for execution by a process. These logged commands are subsequently forwarded to the original shell for execution. Notably, unlike system calls, only executed commands are recorded in this solution.

SystemTap: Since the Strace and Mshell solutions are tailored for specific processes, a comprehensive system-level monitoring approach is necessary when the targeted process is unknown. As shown in Fig. 3c, SystemTap is integrated into the hosted OS to analyze the behavior of all running processes, including the kernel, by recording system calls. This solution is notably less conspicuous to user processes and is better suited for examining modern malware or fileless attacks equipped with sophisticated anti-detection technology.

Intelligent IoT Endpoint Detection and Response (EDR)

With the advantages of high fidelity and scalability, the system-level monitors integrated with the emulated IoT system can capture the runtime behavior of IoT endpoint devices behind the edge, thereby opening the door to enable the detection of malicious behavior. The presence of re-hosted firmware in emulated IoT devices also offers a new potential avenue to implement proactive protection for IoT endpoint devices behind the edge.

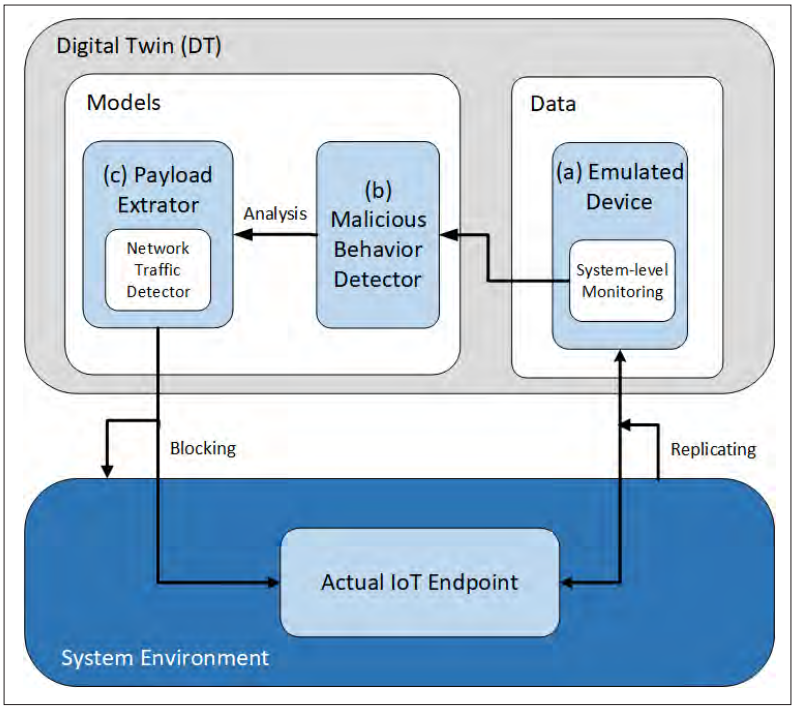


FIGURE 2. Framework for DTs.

Network Architecture and Procedures

Figure 4 describes the network architecture of the proposed intelligent IoT EDR. Different from the abstract DT framework depicted in Fig. 2, this figure concentrates more on the operation procedures and real traffic flow of the EDR . In particular, how the system-level malicious behavior of DT is detected and the corresponding attack is mitigated.

Step 0. Using the full system emulation techniques mentioned earlier, the images of emulated firmware are constructed and stored offline. Once an IoT endpoint device attaches to the EDR platform, its format is analyzed. The corresponding DT is efficiently launched by loading the images into the EDR.

Steps 1. and 2. The inbound traffic to the protected IoT devices mirrors the actual device and the DT. Since the DT is emulated from real firmware, it can be used for responding to the inbound traffic. However, in some cases DT may not send an appropriate response about peripheral devices, actual device assists DT to respond. The integrated system-level monitors intercept all the commands and system calls executed in the DT. In this case, even without malicious binary, the fileless attack can be identified. EDR could leverage the ML algorithm to determine if the downloaded binary or runtime behavior is malicious or not. The details of such endpoint detection will be described next.

Step 3. When malicious actions are detected, the EDR leverages the payload extractor mentioned in Fig. 2c to identify packets containing the malicious payload. Additionally, both the actual device and the DT are rebooted to synchronize their states. Subse-

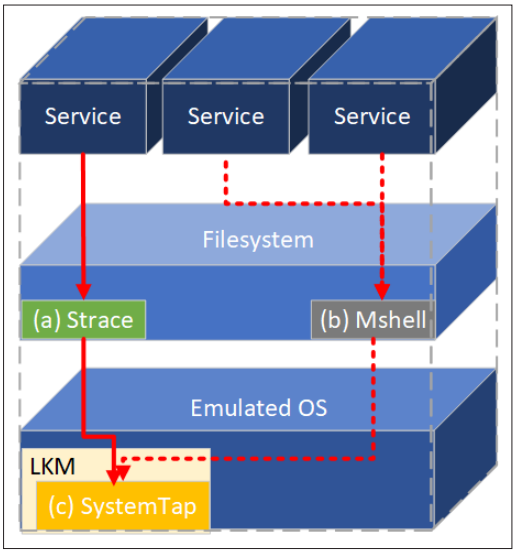


FIGURE 3. System-level monitoring.

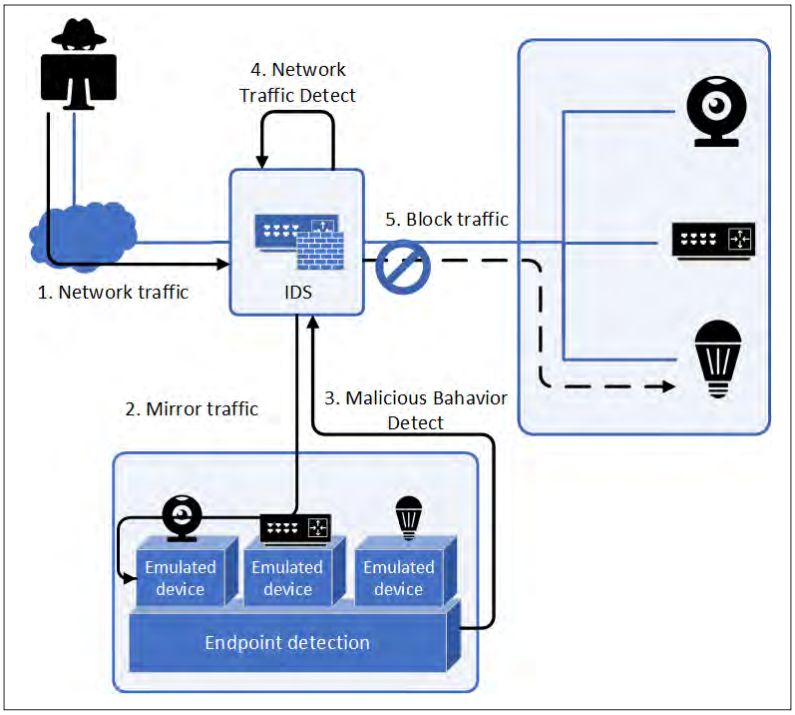


FIGURE 4. Network architecture of intelligent IoT EDR using full system firmware re-hosting.

quently, a YARA rule corresponding to the identified packets is constructed and relayed to the IDS within the EDR, where we look for networking-related command text strings.

Step 4. The IDS inside EDR will update the rule database according to the information received from the DT.

Step 5. The malicious traffic directed to the actual IoT endpoint device is immediately blocked. As a result, endpoint detection and protection are enabled via emulated DT's aid.

Endpoint Detection

Our endpoint detection acquires system-level behavior with the assistance of the emulated DT. The system-level data is then converted into a particular format to serve as input for the proposed detector. This detector consists of three phases:

1. Raw data collection
2. Feature extraction and pre-processing
3. Verification and analysis

We implement an ML-based detector that reasons about the semantics of system-level log sequences for identifying malicious behavior.

Raw Data Collection: The system-level monitor collects the system call sequence as the raw data. The data is labeled according to the type of target process and binary, e.g., malware or benign ware. Following is an example of raw data.

```
734 1629037867.467243 open("/dev/FTWDT101
watchdog", O_RDWR unfinished …
729 1629037867.346495 write(1, "Yowai: Raping
you sorry 0", 24) = 24 0.001728
732 1629037867.664298 read(0, unfinished ...
728 1629485115.479671 close(3) = 0 0.000215
728 1629485115.510892 ioctl(0, TCGETS, 0x7ec-
3da5c) = -1 ENOTTY (Inappropriate ioctl for
device) 0.000199
```

Feature Extraction and PrePprocessing: The system-level monitor collects the system call sequence as the raw data, consisting of system call name, system call parameters, and return value. We simply extract the name of the system call as features for the following processing. The parameters of the system calls are not considered to prevent confusion to the machine learning model. Then we concatenate names of system calls into a chronological sequence. The previous example after feature extraction becomes

```
open write read close ioctl
```

Verification and Analysis: We apply TF-IDF to convert system call sequences into vectors. If the entire dataset contains 167 different system call names, the vector dimension is (1,167). For example, if the dataset only contains the following two system call name sequences:

```
1: read read write write open close ioctl
2: open read write open read write
```

By using TFIDF, the dataset is converted as

```
   close ioctl open read write
1: 0.390548~0.390548~0.277878~0.555756~0.555756
2: 0.000000~0.000000~0.577350~0.577350~0.577350
```

The pre-processed features are then inputted into the ML model for detection, where Support Vector Machine (SVM) and Random Forest (RF) algorithms are applied.

After analyzing the system-level behavior, the packets containing malicious behavior will be captured. To enhance accuracy, we built another ML-based detector that considers the relationship of network traffic for identifying malicious behavior. As shown in Table 1, most IoT endpoint devices test the connection status by constructing DNS queries to a few domains, such as google.com. The cumulative count of DNS queries is beneficial. Additionally, since IoT endpoint devices do not actively establish connections to other devices, the number of unique IP addresses plays an important role in the features. For model selection, we chose the Support Vector Machine (SVM) and Random

Forest (RF) algorithms to train our dataset.

Validation and Performance Evaluation

Experimental Setup

The intelligent EDR in the experimental environment is implemented using an Intel NUC with an Intel Core i3-8109U processor, 32GB RAM, and a 256GB SSD. Ubuntu 18.04 is chosen as the operating system, and Security Onion is selected as the IDS engine with data visualization tools. To avoid direct modification of incoming network packets, Security Onion uses a mirror port to replicate network traffic, so external network interface cards are needed on the NUC to achieve this function.

Regarding the actual IoT endpoint devices for protection, commercial digital video recorders (DVRs), IP cameras, and routers are deployed with ARM, MIPS, and X86 architectures. Depending on the IoT device, we obtain and extract its firmware and activate a virtual DT for the IoT device using firmware emulation techniques, ensuring it possesses the same characteristics as the actual device. At the same time, we compile the corresponding kernel module using a cross-compiler for different architectures, integrate it into the DT, and collect system calls using the system-level monitoring module.

Attack Implementation

Regarding malware attacks, we implemented Mirai and its variants using multiple publicly available proofs of concept (PoCs) and Metasploit modules.[1] Additionally, we implemented four PoCs for fileless attacks targeting endpoint devices, including CVE-2020-10514, CVE-2019-10999, and CVE-2020-10987. Two notorious attacks were selected: buffer overflow and command injection. A buffer overflow occurs when large data is sent, exceeding the buffer size, which can cause system malfunctions or allow attackers to take control. Command Injection is a common type of web injection attack where administrators fail to filter sensitive characters in a website's input form, enabling attackers to send payloads to execute arbitrary commands.

Dataset

By applying the developed malware and fileless attacks mentioned in the previous subsection within our experimental environment, we generated trace results through fuzzing, some of which can be labeled as malware. To avoid a high false alarm rate, we first collect and analyze historical data to understand what typical behavior looks like for commands and log entries. Here are two examples:

1. `` `dd if=/dev/zero of=/dev/sda bs=512 count=1` ``: This command has the potential to cause disk wipe or data corruption. We need to check if this command is executed during routine maintenance or if it is unexpected, which could indicate data corruption.
2. `` `Aug 8 23:45:12 server sshd[1234]: Failed password for invalid user admin from 192.168.1.100 port 22 ssh2` ``: This log entry could indicate a brute force attack or unauthorized access attempt. We compare this against historical failed login attempts to determine if it is part of a larger brute force attack. Moreover, system binaries such as `init`, `/sbin/syslogd`, or `~/bin/sh` were executed to generate datasets labeled as benign. The dataset encompasses three architectures, with the number of traces for ARM, MIPS, and X86 being 457,373; 579,339; and 152,458, respectively.

Type	Feature
Data Size	TCP Upload Bytes, TCP Download Bytes, UCP Upload Bytes, UCP Download Bytes
Packet Count	TCP Upload Packets, TCP Download Packets, UCP Upload Packets, UCP Download Packets
Counter	Total Number of DNS Domain Name, Total Number of Unique IP
Other	Direction of Connection

TABLE 1. Extract feature from network traffic.

Evaluation Matrices

In the evaluation phase, we adopted the common evaluation metrics, namely, accuracy, recall, precision, false-positive rate, and F1-measure, to assess the performance of our proposed method. These metrics are defined based on the following intermediate measures.

- True positive (TP): samples correctly classified as positive.
- False positive (FP): samples incorrectly classified as positive.
- True negative (TN): samples correctly classified as negative.
- False negative (FN): samples incorrectly classified as positive.

Accuracy refers to the proportion of correct judgments of true and false. Precision refers to how much is true when the judgment is true. Recall is the probability of the samples in the positive class being classified correctly:

$$Recall = \frac{TP}{TP + FN}. \quad (1)$$

F1-measure is the weighted average of precision and recall:

$$F1\_measure = \frac{2 * (Recall * Precision)}{Recall + Precision}. \quad (2)$$

Experimental Results

The experiment was conducted on a dataset comprising all samples from our dataset. We utilized 10-fold cross-validation to validate our experimental results, ensuring the robustness of the experiment. The dataset was split into a training dataset, containing seventy percent of the overall dataset, and a testing dataset, containing thirty percent of the overall dataset. The results were averaged over ten independent runs, with the training and test sets determined by 10-fold cross-validation. Table 2 presents the experimental results of the two models. It is evident that Random Forest outperforms the other model, especially in terms of recall (i.e., 100%). Specifically, in the system-level detector, Random Forest achieved an accuracy of 99.22%, an F1-score of 99.29%, precision of 99.05%, and recall of 99.54%, outperforming SVM, which achieved lower accuracy and precision. Similarly, in the network traffic detector, Random Forest maintained a high level of performance, with nearly perfect accuracy

System-level Detector				
Classifier	**Accuracy**	**F1-score**	**Precision**	**Recall**
Random Forest	99.22%	99.29%	99.05%	99.54%
SVM	94.72%	96.78%	96.01%	97.55%
Network Traffic Detector				
Classifier	**Accuracy**	**F1-score**	**Precision**	**Recall**
Random Forest	99.94%	99.8%	99.95%	99.95%
SVM	99.35%	98.47%	99.43%	99.43%

TABLE 2. Performance comparison of Random Forest and SVM in system-level detector and network traffic detector.

(99.94%), F1-score (99.8%), precision (99.95%), and recall (99.95%). These results highlight the superior robustness and reliability of Random Forest in both detection scenarios, making it a more effective choice for IoT EDR systems. In particular, Random Forest's ability to maintain a strong balance between precision and recall suggests that it is highly capable of minimizing false positives and negatives, which is critical in maintaining security without overwhelming administrators with false alerts.

In the experiment, we exploit a command injection vulnerability in a commercial DVR device and then launch the telnet service. After observing the system call \texttt{write} interacting with the web service via a system-level monitor, we identify the packets containing the malicious payload and convert the payload into an IDS rule. For example, due to the command injection vulnerability in the DVR web service, the IDS rule is:

```
alert tcp any any -> any any (msg: "Command
injection"; content: "GET /goform/setUsbUn-
load/.js?deviceName=A;.*"; pcre:"/[a-zA-Z0-9]
{2}/"; sid:101;)
```

Conclusion

In order to enable a feasible detection and response solution for resource-constrained IoT endpoint devices, this article leverages a powerful edge to establish a DT of the actual IoT endpoint devices through firmware emulation. Integrating a system-level monitoring module with a software-defined DT could investigate the precise operational behavior of an IoT device, thereby resolving the drawbacks of typical network-based IDS solutions where only packets can be observed. We propose an ML-based detector in EDR where system calls are leveraged to reason the operational semantics, allowing harmful behavior to be detected. With the aid of IDS, malicious traffic to the actual IoT device can be blocked, thereby achieving endpoint response. Experimental results demonstrate that the EDR successfully captures malware and fileless attacks targeting commercial IoT endpoints deployed behind the edge with high accuracy, reaching 99.94%. As the first IoT DT facilitating the detection and protection of IoT endpoint devices, this article demonstrates the potential of firmware emulation in securing the IoT paradigm. Inspired by this article, many DT applications using firmware emulation are expected to be proposed for the security enhancement of IoT endpoint devices.

[1] https://docs.metasploit.com/docs/modules.html.

Acknowledgments

This work was partly supported by the National Science and Technology Council (NSTC), Taiwan, under Grant 113-2634-F-011-002-MBK.

References

[1] F. Dang *et al.*, "Understanding Fileless Attacks on Linux-Based IoT Devices with HoneyCloud," *Proc. ACM MobiSys 2019*, June 2019, pp. 482–93.
[2] M. Eskandari *et al.*, "Passban IDS: An Intelligent Anomaly Based Intrusion Detection System for IoT Edge Devices," *IEEE Internet Things J.*, vol. 7, no. 8, Aug. 2020, pp. 6882–97.
[3] P. M. S. Sánchez *et al.*, "A Survey on Device Behavior Fingerprinting: Data Sources, Techniques, Application Scenarios, and Datasets," *IEEE Commun. Surveys Tuts.*, vol. 23, 2nd Quarter 2021, pp. 1048–77.
[4] A. Mudgerikar, P. Sharma, and E. Bertino, "Edge-Based Intrusion Detection for IoT Devices," *ACM Trans. Manag. Info. Systems*, vol. 11, no. 4, Oct. 2020.
[5] C. Wright *et al.*, "Challenges in Firmware Re-Hosting, Emulation, and Analysis," *ACM Computing Surveys*, vol. 54, no. 1, Apr. 2021, pp. 1–36.
[6] R. Eramo *et al.*, "Conceptualizing Digital Twins," *IEEE Softw.*, 2021, accepted for publication.
[7] H. Alasmary *et al.*, "SHELLCORE: Automating Malicious IoT Software Detection Using Shell Commands Representation," *IEEE Internet Things J.*, 2021, accepted for publication.
[8] J. Khoury, M. Safaei Pour, and E. Bou-Harb, "A Near Real-Time Scheme for Collecting and Analyzing IoT Malware Artifacts at Scale," *Proc. 17th Int'l. Conf. Availability, Reliability and Security*, 2022, pp. 1–11.
[9] J. Habibi *et al.*, "Heimdall: Mitigating the Internet of Insecure Things," *IEEE Internet Things J.*, vol. 4, no. 4, Aug. 2017, pp. 968–78.
[10] Y. Jia *et al.*, "FlowGuard: An Intelligent Edge Defense Mechanism Against IoT DDoS Attacks," *IEEE Internet Things J.*, vol. 7, no. 10, Oct. 2020, p. 9552–62.
[11] R. Minerva, G. M. Lee, and N. Crespi, "Digital Twin in the IoT Context: A Survey on Technical Features, Scenarios, and Architectural Models," *Proc. IEEE*, vol. 108, no. 10, Oct. 2020, pp. 1785–1824.
[12] H. Elayan, M. Aloqaily, and M. Guizani, "Digital Twin for Intelligent Context-Aware IoT Healthcare Systems," *IEEE Internet Things J.*, vol. 8, no. 23, Dec. 2021, pp. 16,749–57.
[13] G. Mylonas *et al.*, "Digital Twins from Smart Manufacturing to Smart Cities: A Survey," *IEEE Access*, vol. 9, Oct. 2021, pp. 143,222–1,432,492.
[14] M. M. Rathore and S. A. Shah, "The Role of AI, Machine Learning, and Big Data in Digital Twinning: A Systematic Literature Review, Challenges, and Opportunities," *IEEE Access*, vol. 9, Feb. 2021, pp. 32,030–352.
[15] D. D. Chen *et al.*, "Towards Automated Dynamic Analysis for Linux-Based Embedded Firmware," in Proc. NDSS 2016, Feb. 2016.

Biographies

Shin-Ming Cheng (smcheng@mail.ntust.edu.tw) received the B.S. and Ph.D. degrees in computer science and information engineering from the National Taiwan University, Taipei, Taiwan, in 2000 and 2007, respectively. Since 2012, he has been on the faculty of the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, where he is currently a Professor. Since 2022, he has served as the Deputy Director-General in Administration for Cyber Security, Ministry of Digital Affairs. His current interests are mobile network security, IoT system security, malware analysis and AI robustness. He has received IEEE Trustcom 2020 Best Paper Awards.

Yi-Ching Lui (m10815109@mail.ntust.edu.tw) received Master degree in computer science and information engineering from National Taiwan University of Science and Tchnology, Taipei, Taiwan, in 2022.

Nien-Jen Tsai (m11115009@mail.ntust.edu.tw) is an open-source enthusiast who enjoys enhancing computational speed and security. She has contributed to projects such as LLVM, libFuzzer, QEMU, and OpenSSL. Her current research focuses specifically on system and network security.

Bing-Kai Hong (d10815003@mail.ntust.edu.tw) received his B.S. degree in computer science and information engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2018. He is currently a Ph.D. candidate of the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei. He visited EURECOM and NICT Cybersecurity Lab in 2018 and 2019, respectively. His research interests are secure system integration and development using virtualization technologies in mobile networks and IoT systems. He has received a 4-year scholarship of the Ministry of Science and Technology, CISC 2020 and TANET 2021 best paper awards.

Merging Threat Modeling with Threat Hunting for Dynamic Cybersecurity Defense

Boubakr Nour, Sonika Ujjwal, Leyli Karaçay, Zakaria Laaroussi, Utku Gülen, Emrah Tomur, and Makan Pourzandi

Abstract

As technology advances swiftly and the Internet of Things undergoes significant growth, the world is experiencing a surge in data creation. This has resulted in the rapid emergence of novel applications, bringing forth a broader range of intricate and challenging threats that pose difficulties in detection. Therefore, a comprehensive and proactive approach is needed to identify and mitigate security threats. In this article, we combine threat modeling and threat hunting using different approaches in order to provide a more holistic understanding of the security posture of the system, by leveraging the threat model capability in anticipating potential threats and the capability of the threat hunting in identifying evolving and previously unidentified threats. This integration allows for early detection and mitigation of potential threats and enables organizations to enhance their incident response readiness, implement targeted risk mitigation strategies, and fortify their overall cybersecurity posture in the face of evolving and sophisticated threats.

Introduction

In light of the recent advances in digital transformation, organizations are encountering unprecedented opportunities for innovation and growth. This development has the potential to reshape a wide range of industries. For instance, the Internet of Things (IoT), is expected to reach its full potential. Billions of IoT devices will be able to transmit data and interact with centralized systems or each other. The massive expansion of generated data and interconnected devices will significantly enlarge the cyber attack surface, introducing more opportunities for suspicious activities and malicious actors such as Advanced Persistent Threats (APTs) [1]. This escalation is compounded by the diversity of platforms, protocols, and devices, which leads to new attack techniques and threat variants [2].

Some threats may exist in an organization's network for months without being noticed, Mandiant [3] exposed the activities of APT1 (a.k.a. Unit61398) that succeeded in making unauthorized intrusions to over 150 victims including governments and major industries. The APT used sophisticated tools and methods to steal sensitive information and run extensive espionage campaigns to steal hundreds of terabytes of data. Similarly, CrowdStrike [4] revealed that LightBasin (a.k.a. UNC1945) attack has been active since 2016, been unnoticed for 6 years, and was detected in 2022. In such an evolving ecosystem, traditional security measures — which often rely on threat modeling to anticipate vulnerabilities [5] or reactive strategies (i.e., attack detection) to identify existing threats [6], are no longer sufficient. It is essential to empower the predictive power of threat modeling with the proactive capabilities of threat hunting [7].

Security Operation Centers (SOCs) employ a multifaceted approach to cybersecurity, encompassing threat modeling, attack detection, and threat hunting. Threat modeling [8] involves the proactive identification and assessment of potential security threats, enabling organizations to design their systems with robust security from the outset and prioritize their security efforts effectively. This approach not only mitigates risks before they can be exploited but also improves the overall architecture against attacks. In tandem with this, attack detection [9] plays a critical role by monitoring network traffic and system logs to identify any signs of malicious activity. The runtime identification of threats through detection tools minimizes potential damage and helps in understanding the methods used by attackers. In addition to these measures, it is essential to consider threat hunting [10], a proactive defense strategy where security experts delve into various data telemetry to uncover advanced threats that might bypass detection mechanisms. This proactive hunting helps uncover hidden threats, reduces the time attackers spend undetected within the network, and continually enhances security measures and incident response capabilities. Table 1 provides a qualitative comparison between threat modeling, attack detection, and threat hunting. This comparison highlights the differences in focus, methodology, and application among these three critical aspects of cybersecurity. Threat modeling involves anticipating potential threats and vulnerabilities to design robust defenses proactively. Attack detection focuses on identifying and responding to attacks as they occur, emphasizing real-time monitoring and immediate mitigation. Threat hunting, on the other hand, involves proactively searching for and investigating potential threats that have evaded traditional detection mechanisms.

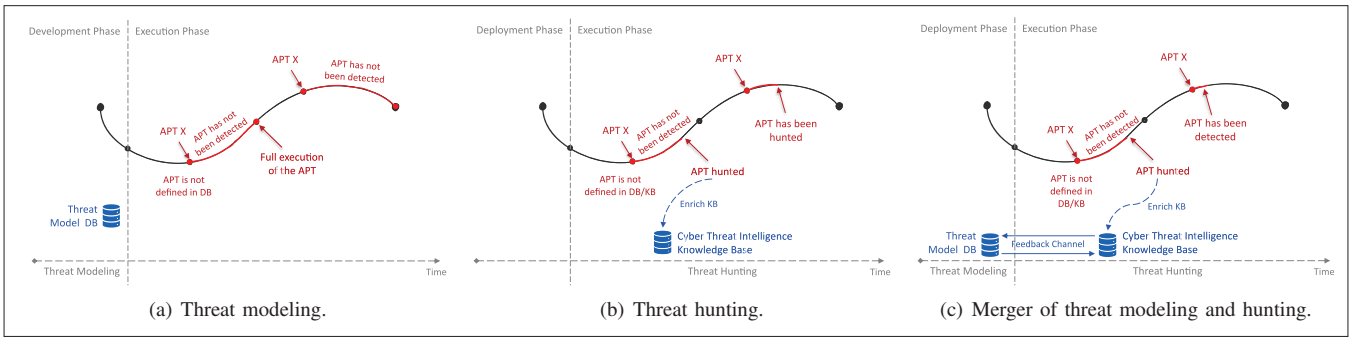
Boubakr Nour and Makan Pourzandi are with Ericsson Security Research, Canada; Sonika Ujjawa and Zakaria Laaroussi are with Ericsson Security Research, Finland; Leyli Karaçay and Utku Gülen are with Ericsson Security Research, Turkey; Emrah Tomur is with Izmir University of Economics, Turkey.

Digital Object Identifier: 10.1109/IOTM.001.2400061

2576-3180/24/$25.00 © 2024 IEEE

	Threat Modeling	Attack Detection	Threat Hunting
Definition	A systematic process to identify and address potential threats during the developmental phase of a system	Continuously monitoring systems to detect and alert on suspicious activities as they happen	A proactive, iterative approach to finding advanced threats that bypass existing detection solutions
Approach	Proactive	Real-Time Monitoring	Proactive
Area	Architecture	Known Threats	Unknown Threats
Execution	During the development phase	Continuous, in real-time	Ad-hoc, based on suspicions or indicators of compromise
Required Expertise	Broad understanding of the system's architecture and potential risk areas	Automated tools and solutions, expertise in configuring and maintaining these systems	Deep knowledge in cybersecurity for analytical and exploratory investigations
Techniques	Data Flow Diagrams, Attack Trees, and STRIDE methodology, etc.	Signature matching, anomaly detection, ML algorithms, etc.	Statistical analysis, Machine Learning, User and Entity Behavior Analytics, etc.
Tools	Microsoft Threat Modeling Tool, OWASP Threat Dragon, etc.	Security Information and Event Management, Intrusion Detection Systems, Endpoint protection platforms, etc.	Advanced threat intelligence platforms, Big data analytics platforms, and custom scripts and algorithms for deep data analysis
Security Capabilities	Preventing threats	Detecting known threats and generating alerts for immediate response	Proactively discovering new, unknown threats, and generating threat intelligence
Outcome	Preventive measures for identified threats	Immediate response to threats and alerts	Enhanced security posture by proactive threat identification

TABLE 1. Qualitative comparison between threat modeling, attack detection, and threat hunting.



(a) Threat modeling. (b) Threat hunting. (c) Merger of threat modeling and hunting.

FIGURE 1. Illustrative example of the usage of threat modeling, threat hunting, and merger of threat modeling and hunting to proactively detect advanced threats.

Motivation Example: To effectively hunt and detect evolving cyber threats, it is crucial to combine the proactive foresight of threat modeling with threat hunting capabilities.

Figure 1a illustrates a timeline-based scenario of detecting a sophisticated attack (e.g., APT X) that has not been defined in the knowledge base nor triggered known Indicator of Compromises (IoC) during its execution. Given that APT X is not defined in the threat model database — as threat modeling is executed during the development phase and may not contain a definition/signature of APT X, the detection mechanism will not be able to detect this attack and the attack will be fully executed. Similarly, another APT occurrence might not be detected in the future either. This is where threat hunting becomes invaluable by proactively searching for novel IoC and unconventional attack patterns. Yet — as shown in Fig. 1b, the threat hunter might take time and effort to manually hunt a threat using different tools and security knowledge and uncover its pattern before enriching the Cyber Threat Intelligence (CTI) Knowledge Base. However, due to the attack's stealthiness, another occurrence of the attack might be identified but not immediately. However, a synergistic approach that combines threat modeling with threat hunting by updating the threat model database in run-time will enhance attack detection capabilities. As illustrated in Fig. 1c, the threat model database will be enriched by the information collected by the threat hunter which will lead to faster detection of new attacks in the future.

The integration of threat modeling and threat hunting, therefore, helps in updating the threat model at the run-time with the newly discovered threats, while guiding the threat hunter with the most prioritized threats. This integration forms a comprehensive defense strategy against known threats while anticipating and adapting to new and evolving cybersecurity challenges, thereby strengthening the overall defensive posture. The approach illustrated in the motivating example is one of the three proposed integration approaches, which will be elaborated more later. This approach necessitates the effective integration of real-time data from threat hunting into traditionally more static threat models, requiring robust systems capable of processing and adapting to live threat information. In addition, establishing a systematic feedback loop is essential, where threat hunting continuously informs and updates threat models, ensuring they remain relevant and adaptable to emerging threats.

Therefore, this article presents a multifaceted exploration of the cybersecurity landscape, with a

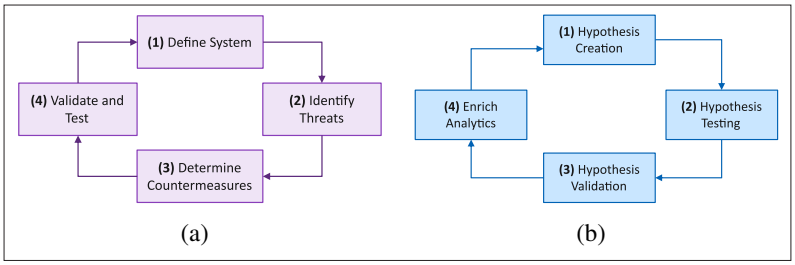


FIGURE 2. Steps of (a) threat modeling, and (b) threat hunting.

particular focus on guiding the threat hunting with prioritized threats from the threat model and updating threat modeling in run-time using threat hunting. First, we conduct a thorough review of threat modeling, delineating its critical aspects and inherent challenges. Second, we provide an in-depth analysis of threat hunting, where we examine its intricacies and the difficulties encountered in its implementation. Finally, we discuss the integration of threat modeling with threat hunting, offering a comparative analysis of various approaches. This comparison not only highlights the advantages and disadvantages of each approach but also opens avenues for future research directions. By presenting this synthesis, the article aims to contribute to the understanding and advancement of proactive cybersecurity strategies.

Threat Modeling

Threat modeling [8] is a set of techniques (e.g., Data Flow Diagram — DFDs) used to identify and assess possible security risks and vulnerabilities in a system. Using these techniques, it is convenient to recognize the threats and proactively mitigate them to improve the security posture of the system.

Threat Modeling Steps

Threat modeling involves four main steps, as illustrated in Fig. 2a:

1. *Define The System:* This step involves key actions to define the scope, architecture, components, and assets of the system. In order to highlight areas of security concerns, DFDs are used to indicate how data moves within the system and eventually define the trust boundaries.
2. *Identify Threats:* This step involves identifying threat actors and potential threats and vulnerabilities in the system. It is beneficial to specify which security properties such as confidentiality, integrity, and availability, are affected by the identified threats. Various threat modeling methodologies can be used to guide the threat identification process.
3. *Determine Counter Measures:* This step focuses on identifying and selecting appropriate measures to mitigate or address the identified threats.
4. *Validate and Test:* This step verifies the effectiveness of the selected security measures to mitigate the identified threats which may include penetration testing, code review, security controls, etc.

Threat Modeling Methodology

There are different methodologies and tools available for conducting threat modeling, including Microsoft's STRIDE, PASTA, and OCTAVE, to cite a few. The decision on which methodology to pursue depends on the requirements and context of the system. In addition, several approaches to threat modeling are often used in the literature including asset-centric, data-centric, threat-centric, and system-centric, each with its focus and methodologies [11, 12]. Assets can include sensitive data, intellectual property, hardware components, or even software components. This approach involves understanding the flow of assets through the system, identifying potential vulnerabilities that could compromise asset security, and assessing the impact of potential threats on these assets. Data-centric threat modeling emphasizes the protection of data and its associated data flows within a system. It involves identifying critical data, understanding its lifecycle, and assessing how data is stored, processed, transmitted, and accessed. Threats are then analyzed in terms of their potential to compromise the confidentiality, integrity, or availability of the data. Threat-centric threat modeling focuses on identifying potential threats and vulnerabilities first and then assessing their impact on the system. This approach begins with a list of threats, which can be drawn from common threat databases or industry-specific knowledge. Each threat is then examined to understand its potential consequences and the specific assets or components it could affect. System-centric threat modeling takes a holistic view of the entire system or application architecture. It involves examining the system's components, interactions, and dependencies to identify vulnerabilities and potential weaknesses. This approach considers the overall system design, including trust boundaries.

Threat Modeling Challenges

Despite the importance of threat modeling, several challenges can hinder its effectiveness:

- *Expert dependency:* Threat Modeling is highly dependent on expert knowledge and expertise both in the system and security domains. This dependency emphasizes the need for continuous training and upskilling of security professionals to stay abreast of evolving technologies and emerging threats.
- *Resource Constraints:* Performing thorough threat modeling requires time, expertise, and computational resources.

  This lack of proper resources can be particularly detrimental. Although automated tools and techniques can be used, they cannot entirely replace the nuanced understanding of security professionals.
- *Assessment Subjectivity:* Assessing the impact and likelihood of threats involves a degree of subjectivity. Different stakeholders and domains may have varying perspectives on the severity of particular threats. It is crucial to establish and agree on a set of criteria for threat assessment across the organization. This ensures a more uniform approach to prioritizing threats.
- *Dynamic Threats:* Unexpected threats not initially anticipated may be introduced by new features and technologies that can quickly outdate threat models. Keeping threat models up-to-date in dynamic environments, especially in agile development settings, can be a persistent challenge.
- *Unknown Vulnerabilities:* Unknown vulnerabilities or system behaviors may be missed since threat modeling frequently depends on existing knowledge about the system. Engag-

ing in activities like penetration testing and red team exercises can help uncover these hidden vulnerabilities.

Threat Hunting

Threat hunting [10] is a proactive cybersecurity practice used to continuously search for indicators of compromise or any malicious activity that might evade the conventional detection system. Unlike traditional security measures [13] — which rely on automated alerts and are inherently reactive, threat hunting seeks to identify threats before they escalate and are fully materialized.

Threat Hunting Steps

The expertise and qualification of the security expert (i.e., threat hunter) play a vital role in conducting successful threat hunting [7]. In doing so, the hunter follows four main steps, as illustrated in Fig. 2b:

- *Hypothesis Generation*: Based on current threat intelligence, known vulnerabilities, and observed activities, the hunter develops assumptions about existing potential threats.
- *Hypothesis Testing*: For each generated hypothesis, the hunter examines the collected data using various tools, techniques, and expertise to verify or refute the existence of threats.
- *Hypothesis Validation*: In order to validate the tested hypotheses, the hunter investigates the nature and scope of validated hypotheses, as well as the potential impact to identify attack patterns and assign the threat to a threat actor class.
- *Enrich Analytics*: The hunter documents the findings from the validation step in CTI knowledge base as shown in Fig. 1b, in order to improve attack detection capabilities and refine the hunting process.

Threat Hunting Methodology

Threat hunting involves a proactive, iterative approach to hunting advanced threats that elude detection mechanisms. This methodology starts with developing hypotheses for potential attacks based on current threat intelligence and a deep understanding of the organization's environment [14]. Myriad and tedious data telemetry — from logs, network traffic, and endpoints, are collected and deeply scrutinized, often using existing security tools and manually based on the hunter's expertise [15]. Integration of threat intelligence — such as MITRE ATT&CK framework and collaborative CTI knowledge base, is essential to perform effective hunting. The process involves searching for signs of compromise and refining hypotheses based on findings. Once a threat is identified, it is confirmed and contained, followed by thorough documentation and reporting. Insights gained are used to enhance security measures and inform future hunts, emphasizing continuous improvement and adaptation to the evolving threat landscape.

The methodology allows organizations to enhance detection and response to threats proactively. Involvement of an incident response team ensures that findings from threat hunting are effectively acted upon, therefore leading to containing and mitigating identified threats. Such proactive security practices carry different benefits to the SOC, such as early detection of hidden threats, leading to quicker and more effective response strategies, and reducing the potential impact of breaches. Indeed, threat hunting fosters a deep understanding of adversary tactics, which enhances the organization's security measures and enables the development of tailored defenses. This proactive approach not only helps in reducing the attack surface by identifying suspicious and malicious activities but also builds a comprehensive threat intelligence knowledge base.

Threat Hunting Challenges

Regardless of the aforementioned benefits, threat hunting presents significant challenges:

- *Extensive Experience and Expertise*: Threat hunting requires a deep understanding of both the system/network architecture and the potential threats. This expertise is not easily acquired and involves knowledge of various domains like network protocols, system vulnerabilities, and the latest malware tactics.
- *Managing the Surge in Event Volume*: As organizations expand and the number of connected devices increases, the volume of events (logs, alerts, etc.) that need to be monitored grows exponentially. This deluge of data makes it challenging for threat hunters to identify genuine threats. The sheer volume can overwhelm systems and analysts, leading to alarm fatigue, delayed responses, and potentially missed threats.
- *Navigating the Risk of False Detection*: In the process of detecting threats, systems often flag benign activities as malicious (i.e., false positive) or stealthy activities as benign (i.e., true negative). A high rate of false positives can waste valuable time and resources, as security teams must investigate each alert while a high rate of true negatives will leave the threat lurking in the organization unnoticeable.
- *Manual Intervention*: Many aspects of threat hunting still require manual intervention, such as generating hypotheses, analyzing patterns, correlating events, and investigating the hypotheses. Manual processes are time-consuming and prone to human error. They also do not scale well in larger environments, making it difficult to keep pace with the rapid evolution of cyber threats.
- *Continuous Race Against Evolving Threats*: Cyber threats are continuously evolving, with attackers constantly developing new techniques and tools to bypass security measures. Staying informed about the latest threat landscape is a significant challenge. It requires continuous learning and adaptation of security strategies, which can be resource-intensive and difficult to manage alongside other responsibilities.

Combining Threat Modeling with Threat Hunting

Combining threat modeling with threat hunting yields a comprehensive understanding of the full lifecycle of threats, spanning from conceptual design vulnerabilities to their materialization in the real world. In fact, through threat modeling, organizations can effectively prioritize their hunting efforts, concentrating on high-risk areas for more efficient resource deployment. This strategy is inherently adaptable and remains dynamic, allowing the threat model to be updated with new findings from threat

Threat hunting involves a proactive, iterative approach to hunting advanced threats that elude detection mechanisms.

Using threat modeling, it is possible to prioritize threats based on risk assessment results and allocate resources more effectively for threat hunting.

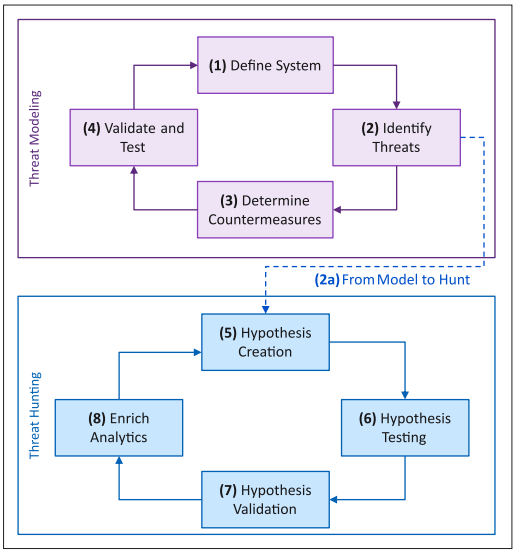


FIGURE 3. From threat modeling to hunting.

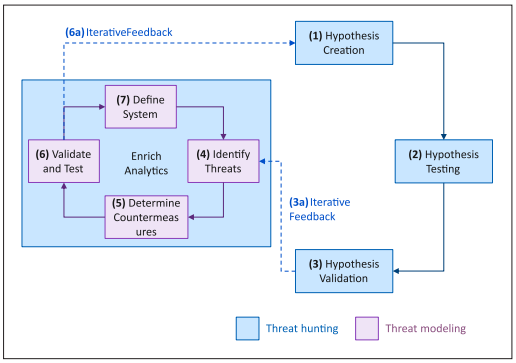


FIGURE 4. Iterative hunting-modeling feedback.

hunting, which in turn, informs adjustments in hunting strategies. The merger of these methodologies significantly reduces the attack surface by identifying potential vulnerabilities and weak points during the design phase and maintaining an ongoing search for operational compromises. Furthermore, when threats are detected during hunting, the pre-existing threat model database provides vital context, enabling quicker and more precise remediation efforts, as illustrated in Fig. 1c and thereby enhancing the organization's overall defensive response. In the following, we present three combination approaches to use threat modeling and threat hunting in run-time.

From Threat Modeling to Hunting

The process of threat hunting often begins with a set of intelligence inputs that guide the hunter. This combination approach aims to use threat modeling to guide and prioritize threat hunting activities. As it is illustrated in Fig. 3, threat modeling, mainly step (1) and step (2), becomes an initial step for threat hunting where key assets and potential threats are identified. This approach enables hunting activities to focus on system assets that are deemed highly valuable, and proactively search for signs of compromise in those assets. It is useful to define IoCs that are tailored to the characteristics and behavior associated with the threats outlined in the threat model. In addition, threat scenarios that consider various factors, such as threat actors, their objectives, and the likely paths they would take to achieve their goals, can be used as a blueprint to drive a threat hunting campaign. Thus, threat hunters can continuously look for novel IoCs based on those associated with the defined threat scenarios instead of waiting for automated security tools to raise alerts. The threat hunting campaigns can be structured either based on time or the severity and likelihood of the threat as defined in the threat model.

Approach Prospects: Using this combination approach allows for efficient and focused threat hunting, driven by the threat model that contains clear information on important threats an organization is looking for. Using threat modeling, it is possible to prioritize threats based on risk assessment results and allocate resources more effectively for threat hunting. However, this approach might miss identifying some important and new threats uncovered during the threat hunting process if the threat model is not comprehensive. In addition to that, threat models are often static and do not easily accommodate new or emerging threats, making it difficult to stay up-to-date with the rapidly changing cybersecurity environment and leading to outdated hunting directives.

Iterative Hunting-Modeling Feedback

This approach starts with threat hunting by using the threat model database as the basis for an initial round of hunting. As illustrated in Fig. 4, the outputs from the hunting phase are used to update and improve the threat model database (steps (4)–(7) which is considered as one step in threat hunting — enrich analytics), and the latter is used to guide the hunting step (6a). The uncovered threats and anomalies by the threat hunting are used to update and refine the threat model database (i.e., adding new threats, adjusting the risk scores of existing threats, or revising countermeasures) through a continuous improvement channel. In fact, the threat modeling process is triggered to identify the involved threats, define any countermeasures, and validate and test them. The system definition might also be updated based on the identified threat by threat hunting. The loop is closed by going back to the hunting. This iterative process ensures that both the model and the hunting strategies are continuously refined and are up-to-date.

Approach Prospects: This approach allows for a dynamic adaptation to new threats and changes in the organization environment as the threat modeling database is continuously improved by the new threats identified by the threat hunting. Indeed, the iterative nature ensures that both the model and the hunting strategies are continuously refined through a feedback mechanism where findings from each approach can help refine the other, enhancing the overall effectiveness. The continuous iterations can help in fine-tuning the detection mechanisms, thereby reducing false positives over time. Yet, managing the feedback loop and constant updates can become complex and challenging, requiring significant time and human resources to be effective. In addition, handling large amounts of data generated by constant iterations might be oppressive and require intensive resources.

Common Hunting-Modeling Tooling

This approach allows performing threat modeling and threat hunting activities on the same platform using a tool or a set of integrated tools (i.e., hybrid solution). The tool should be a specialized software capable of identifying threats in the current

system while proactively creating new hypotheses for threat hunting purposes. As illustrated in Fig. 5, this approach results in identifying threats for threat modeling while generating new hypotheses for threat hunting activity. To this purpose, both threat modeling and hunting activities should ideally be executed in a hybrid mode, where they complement and inform each other, and hence have access to up-to-date CTI databases, system data, incident alerts, and other logs repositories. As a result of common hunting-modeling activity, the tool should have the capability to effectively populate the threat model database as well as threat intelligence databases. This approach should support features of both activities, e.g., threat modeling visualization, countermeasures suggestion and generation, and hunting analytics, to perform a seamless operation of the two activities. By operating in tandem, threat modeling identifies potential vulnerabilities while threat hunting proactively searches for indicators that these vulnerabilities are being exploited. This synergy leads to a robust defense against new vulnerabilities and threats.

Approach Prospects: This approach allows the operations involved in both activities to be streamlined due to the use of a single tool, i.e., once a hypothesis is validated as shown in step (3a), the same tool can generate countermeasures corresponding to that threat, as shown in step (4a). Also, due to the tight integration of modeling and hunting activity enabled by this approach, the data used in both activities can be up-to-date and consistent with each other. In addition, centralized reporting and analytics make it easier to track threats and measure the performance of both activities. However, the operational logic used in the approach for the integration of both activities can be complicated and would require meticulous expertise in threat modeling, hunting, and domain knowledge. It is paramount to ensure the correct and efficient execution of integrated operations without sacrificing the accuracy and timely detection of threats from both activities.

Discussion

Combining threat modeling and hunting enhances an organization's ability to detect and respond to threats in a more proactive and informed manner, leveraging the strengths of both approaches to build a resilient cybersecurity posture.

The first two approaches attempt to empower one activity using the output of the other while running both activities separately. The feedback mechanism between activities provides a nudge in the right direction. These approaches enable security analysts to optimize any operation of any activity without worrying about its effect on the other activity. In general, efficient execution of one activity would only propel the other activity in the right direction. These two approaches can be leveraged in a scenario where threat modeling and threat hunting activities are running or need to run in parallel without many inter-dependencies so that they can be optimized and augmented individually whenever required. The third approach boasts of integrating the two activities warranted by a hybrid process. While this approach promises to provide a unified view of this integration, it instills logical complexities related to the intertwining of operations of threat modeling and threat hunting activity. The resulting tool might make it complex to refine an operation due to its additional dependencies on the operation of the other activity. This approach is comparatively complicated and needs

- Correct and efficient intertwining of operations of both the activities
- Access to the relevant data and databases to perform the intertwined operations of these activities efficiently.

Regardless of the used approach, organizations can balance resource allocation by leveraging automated tools to supplement expert analysis, prioritizing critical threats to optimize the use of limited resources, and fostering cross-disciplinary training to broaden the skill sets within existing teams. Organizations can more effectively integrate threat hunting and threat modeling, enhancing their overall security posture. In addition to that, the models are constantly updated with new threat intelligence by industrial and governmental partners and augmented/ integrated with other knowledge bases. This intelligence stems from internal as well as external security reports and assessments, and the discovery of new patterns.

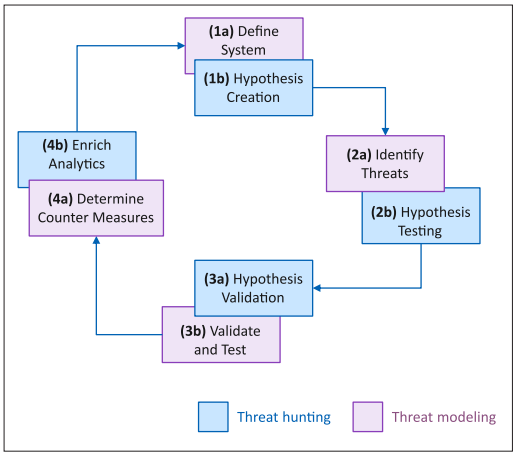


FIGURE 5. Common hunting-modeling tooling.

This approach should support features of both activities, e.g., threat modeling visualization, countermeasures suggestion and generation, and hunting analytics, to perform a seamless operation of the two activities.

Challenges and Research Directions

Combining threat hunting with threat modeling to enhance attack detection and strengthen the proactive defense lines is a promising approach in cybersecurity. However, it does present unique challenges and opens various potential research directions:

Challenges

- *Integration Complexity*: Merging threat hunting and threat modeling requires seamless integration of two distinct methodologies. This integration can be complex (i.e., in terms of involved methodologies), and can involve the alignment of threat hunting and threat modeling process (i.e., which process should be triggered first) and data flow (i.e., what data and information need to be circulated).
- *Scalability and Efficiency*: Ensuring that the combined solution scales effectively in large and dynamic environments while maintaining efficiency is a significant challenge. This involves handling vast amounts of data, numerous potential threat scenarios, and data flow between hunting and modeling components.

- *Real-Time Analysis*: The ability to analyze threats in real-time and respond promptly is crucial. This requires advanced analytics capabilities and the potential use of automated processes relying on machine learning. The latter can be difficult to implement effectively due to the nature of data and stealthy attacks.
- *Accuracy of Threat Models*: The effectiveness of threat modeling depends on the accuracy and comprehensiveness of threat hunting, while the efficiency of threat hunting depends on the quality of defined threats in threat modeling. Keeping these models up-to-date with evolving threat landscapes is a continuous challenge.

Potential Research Directions

- *Automated Integration Techniques*: One of the major challenges would be how to ensure the systematic integration of both activities through a common tool without sacrificing the efficiency/benefits of each activity. Profound expertise will be needed for the intertwining of operations, e.g., how system definition would assist and improve threat hypotheses generation and vice versa. These intertwinings would affect the functioning and performance of the next stage of operations.
- *AI-Driven Models for Real-Time Analysis*: Developing advanced AI algorithms and models for real-time threat analysis within the combined approach can help in learning and adapting to evolved threats. Empowering this with continuous data streams will help in improving security strategies and postures based on the provided feedback between modeling and hunting activities.
- *Dynamic Threat Modeling*: New ways to dynamically update threat models in response to emerging threats identified by threat hunting can enhance the solution's effectiveness. Crowdsourced threat intelligence can be integrated to keep the threat models updated, yet a comprehensive understanding of attack infrastructure is mandatory to not impact the underlying business.
- *Reducing False Positives/Negatives*: With the massive number of events received per second, developing sophisticated analytics techniques to minimize the rate of false positives and negatives in attack detection is very crucial. Implementing automated AI models and semantic-driven correlation techniques can help differentiate between normal and anomalous behaviors more accurately. Yet, those models will not replace the security experts but will assist them in enhanced decision-making.

Conclusion

The symbiotic merger of threat modeling and threat hunting activities is expected to enhance the security posture of a system due to the merging of strategic foresight provided by the threat modeling and tactical agility incorporated by the threat hunting activity. In this work, using an illustrative example, we motivated how a harmonized merger between the two activities paves the way to faster and more proactive detection of future threats in the system. We proposed three different approaches to merging the two activities along with the advantages and disadvantages of each approach. We also presented the challenges impelled by the merger and the potential research directions. As already stated, ensuring the efficient integration of activities, and dynamic updating of knowledge base and threat model database as a result of execution of these intertwined activities will be an active challenge. In addition, AI-driven models for real-time data analysis with accurate attack detection would further assist in the efficient integration of these activities.

Acknowledgment

This work was supported by Scientific and Technological Research Council of Turkey (TUBITAK) through the 1515 Frontier Research and Development Laboratories Support Program under Project 5169902, and has been partly funded by the European Commission through the Horizon Europe/JU SNS project Hexa-X-II (Grant Agreement No. 101095759).

References

[1] J. Khoury *et al.*, "Jbeil: Temporal Graph-Based Inductive Learning to Infer Lateral Movement in Evolving Enterprise Networks," *IEEE Symp. Security and Privacy (SP)*, IEEE Computer Society, 2023, pp. 9–9.
[2] Google Cybersecurity Action Team, "Threat Horizons — April 2023 Threat Horizons Report;" https://services.google.com/fh/files/blogs/gcat_threathorizons_full_apr2023.pdf, 2023.
[3] "APT1: Exposing One of China's Cyber Espionage Units," Mandiant, Tech. Rep., 2013.
[4] CrowdStrike, "LightBasin: A Roaming Threat to Telecommunications Companies;" https://www.crowdstrike.com/blog/an-analysis-of-lightbasin-telecommunications-attacks/, 2021, accessed Feb 2023.
[5] H. Binyamini *et al.*, "An Automated, End-to-End Framework for Modeling Attacks from Vulnerability Descriptions," arXiv, 2020.
[6] A. Alsaheel *et al.*, "ATLAS: A Sequence-Based Learning Approach for Attack Investigation," *USENIX Security*, 2021.
[7] Sqrrl Data, Inc., "A Framework for Cyber Threat Hunting;" https://www.threathunting.net/files/framework-for-threat-hunting-whitepaper.pdf, 2018.
[8] W. Xiong *et al.*, "Threat Modeling – A Systematic Literature Review," *Computers & Security*, 2019.
[9] Z. Li *et al.*, "Threat Detection and Investigation with System-Level Provenance Graphs: A Survey," Computers & Security, 2021.
[10] B. Nour *et al.*, "A Survey on Threat Hunting in Enterprise Networks," *IEEE COMST*, 2023.
[11] M. Tatam *et al.*, "A Review of Threat Modelling Approaches for APT-Style Attacks," Heliyon, 2021.
[12] M. Muckin *et al.*, "A Threat-Driven Approach to Cyber Security," Lockheed Martin Corporation, 2014.
[13] S. M. Milajerdi *et al.*, "HOLMES: Real-Time APT Detection through Correlation of Suspicious Information Flows," *IEEE S&P*, 2019.
[14] B. Nour *et al.*, "Automa: Automated Generation of Attack Hypotheses and Their Variants for Threat Hunting using Knowledge Discovery," *IEEE Trans. Network and Service Management*, 2024.
[15] B. Nour *et al.*, "Accurify: Automated New Testflows Generation for Attack Variants in Threat Hunting," *FPS*, 2023.

Biographies

Boubakr Nour (boubakr.nour@ericsson.com) is an experienced researcher at Ericsson, Montréal, Canada.

Sonika Ujjwal (sonika.a.ujjwal@ericsson.com) is a security researcher at Ericsson, Jorvas, Finland.

Leyli Karaçay (leyli.karacay@ericsson.com) is a senior researcher at Ericsson, Istanbul, Turkey.

Zakaria Laaroussi (zakaria.laaroussi@ericsson.com) is a senior researcher at Ericsson, Jorvas, Finland.

Utku Gülen (utku.gulen@ericsson.com) is a senior researcher at Ericsson, Istanbul, Turkey.

Emrah Tomur (emrah.tomur@izmirekonomi.edu.tr) is an adjunct lecturer at Izmir University of Economics, Izmir, Turkey. He was with Ericsson Security Research, Istanbul, Turkey.

Makan Pourzandi (makan.pourzandi@ericsson.com) is a research leader at Ericsson, Montréal, Canada.

IEEE
INTERNET OF THINGS
MAGAZINE

CALL FOR PAPERS

APPLICATIONS OF LARGE LANGUAGE MODELS IN INTERNET OF THINGS

BACKGROUND

Large Language Models (LLMs) with their advanced natural language processing capabilities, offer numerous benefits, such as improved user interaction, enhanced data analysis, and context-aware services. The integration of LLMs and 6G telecom networks into the Internet of Things (IoT) represents a significant advancement in enhancing the intelligence and interactivity of IoT systems. The LLMs can understand and process complex human language, making IoT devices more user-friendly and capable of responding intelligently to voice commands and text inputs. The applications of LLMs to IoT are vast and varied. LLMs can be employed in smart homes for voice-activated control and personalized user experiences, in industrial IoT for predictive maintenance, anomaly detection and report analysis, and in healthcare IoT for patient monitoring and real-time data analysis. Additionally, LLMs can enhance customer service through intelligent chatbots and virtual assistants, providing timely and contextually relevant information to users. As IoT continues to evolve, the integration of LLMs will play a crucial role in driving innovation and delivering smarter, more responsive, and efficient IoT solutions.

Integrating LLMs into IoT involves several challenges, including the significant computational power and energy consumption required by LLMs, which can strain the limited resources of IoT devices. Real-time processing demands pose latency issues, especially when balancing edge and cloud computing. Ensuring data privacy and security is critical, given the sensitive nature of IoT data and the vulnerability of LLMs to adversarial attacks. Scalability in deployment and maintenance, along with ensuring compatibility and integration across diverse IoT platforms adds complexity. The high costs of implementation and ongoing operations further complicate the integration. Additionally, this Special Issue (SI) aims to optimize model efficiency without sacrificing performance, ensure reliable and accurate responses, and design user-friendly interfaces, which are essential for a positive user experience. Addressing these challenges requires advancements in AI, hardware, software, and cybersecurity to unlock the full potential of LLM-enhanced IoT systems.

Topics of Interest

Topics of primary interest include, but are not limited to, the following scopes:

- Optimizations for reducing the computational and memory footprint of LLMs training, fine-tuning, and inference in resource-constrained IoT environment.
- Optimizations for deploying LLMs on edge-cloud devices to reduce inference latency and improve real-time processing capabilities in IoT applications.
- Optimizations for reducing the energy consumption of LLMs in IoT environment.
- Optimizations of LLM systems and algorithms to meet the real-time processing requirements of IoT applications.
- Efficient and low-cost hardware design to accelerate LLM computation in IoT systems.
- Principles and user interface design to provide a seamless and intuitive user experience.
- Privacy and Security for safeguarding sensitive IoT data and ensuring the secure deployment and operation of LLMs.
- Standards and protocols to ensure seamless integration and interoperability/portability of LLMs with diverse IoT platforms and devices.
- Practical testbeds, datasets, and use cases of integrating LLMs in IoT applications.
- Wide-scale experimentation and performance scalability of LLM-enhanced IoT solutions.

SUBMISSION GUIDELINES

Manuscripts should conform to the *IEEE Internet of Things Magazine* standard format as indicated in the Information for Authors section of the Article Submission Guidelines.

All manuscripts to be considered for publication must be submitted by the deadline through the magazine's IEEE Author Portal site. Select the appropriate issue date and topic from the "Please Select an Article Type" drop-down menu.

IMPORTANT DATES

Submission Deadline
30 November 2024

Decision Notification
10 April 2025

Final Manuscript Due
20 May 2025

Publication Date
Third Quarter 2025

GUEST EDITORS

Gang Sun (Lead Guest Editor)
University of Electronic Science and Technology of China, China
gangsun@uestc.edu.cn

Dusit Niyato
Nanyang Technological University, Singapore
DNIYATO@ntu.edu.sg

Jiacheng Wang
Nanyang Technological University, Singapore
jiacheng.wang@ntu.edu.sg

Nelson Fonseca
University of Campinas, Brazil
nfonseca@ic.unicamp.br

Paolo Bellavista
University of Bologna, Italy
paolo.bellavista@unibo.it

Shu-ping Yeh
Research Lab at Intel, USA
shu-ping.yeh@intel.com

Scenario Co-Design for Systemic Evaluation of Connected and Automated Mobility Setups

Manon Eskenazi, Fabien Kaptue Bopda, Mwendwa Kiko, Natalia Kotelnikova-Weiler, and Daphne Tuncer

Abstract

Connected and automated vehicles are today at various stages of development. Due to their transformative potential, both on the existing transport system and on urban spaces, it is essential to investigate their impacts from a systemic perspective. A multi-factor evaluation cannot be based only on experimental setups. Projections of up-scaled, operational connected and automated mobility (CAM) services are required. In this article we propose TRESSY, a scenario-building approach for CAM service up-scaling. TRESSY follows a four-step model that aims to generate mid-term projections of relevant services based on foreseeable technological and infrastructure developments. We applied TRESSY in a multi-stakeholder CAM pilot project where experts collaborated on the design of a range of up-scale service scenarios and their associated technical systems and infrastructures. The results obtained show how TRESSY can facilitate the collaboration between heterogeneous stakeholders working on representative niche, critical technical systems of future mobility.

Introduction

Increasing levels of vehicle's connectivity and automation have been paving the way to the development of mobility services that take advantage of automated driving systems to operate. Connected and automated mobility (CAM) services build on the provisioning of converged, critical Internet-of-Things (IoT)-enabled infrastructures for the communication, supervision and electrification of vehicle fleets, that involve a various set of stakeholders. The ability to anticipate the systemic impacts of these services, i.e., in terms of security, environmental footprint, socio-economics, governance, for the different stakeholders is essential to inform future CAM deployments in various contexts. To conduct an integrated assessment of these services contributes to gaining knowledge on the application of the connected and automation technology to the mobility domain that goes beyond prototyping and simulation. It helps demonstrate how to plan, deliver and exploit services that require specific infrastructures and operations, to local authorities and policy-makers in particular.

It is today challenging to evaluate the impact of CAM services from a systemic perspective. Despite recent examples of early deployments, notably in the US[1] and in China,[2] the number of large scale experiments or operational services that can serve as a basis for a general evaluation is limited. Impact assessment studies commonly rely on small scale localized field trials or are carried out in simulated environments, which constrains the scope of the factors to be evaluated. In addition, CAM is often considered as a combination of separate systems, i.e., the vehicle, different types of infrastructures, IoT equipment, that are in general evaluated independently and hinders the real effects of the deployment of such services.

In this article we present TRESSY (Time, REgion, Service, SYstem), an approach that addresses the limitations of existing evaluation strategies by enabling inter-stakeholder and cross-evaluation of CAM setups through the co-construction of scenarios that define realistic up-scale services for a set context and specify their associated infrastructures. While several methods have been proposed in the literature to build coherent scenarios of future technology and service deployments in various domains [1, 2], including in the transport domain [4, 5], these are not well-suited for the case of a stand-alone, innovative technology characterized by converged infrastructures. In particular, existing solutions mainly adopt a *top-down* approach that assumes generalized penetration of the technology, which is not adapted to the high uncertainties associated with the long term deployment of CAM services. In contrast, the development of scenarios necessitates an approach that

1. Accommodates flexible setup options
2. Is fine-grained enough to capture all the aspects of a mobility service evaluation
3. Can facilitate the collaboration between stakeholders with heterogeneous perspectives and constraints in the definition of the envisioned scenarios.

TRESSY addresses these requirements by following a *bottom-up* strategy that develops scenarios from a specified context. In particular, TRESSY decomposes the design of a scenario in different steps in order to define, identify and specify for a set time horizon and a set deployment region, the characteristics of a CAM service, including the vehicle, the supervision system, the communication and road infrastructures, as well as

[1] https://www.nytimes.com/2023/10/24/technology/cruise-driverless-san-francisco-suspended.html; accessed 05-08-24.

[2] https://www.apollo.auto/; accessed 05-08-24.

Manon Eskenazi, Fabien Kaptue Bopda, Natalia Kotelnikova-Weiler, and Daphne Tuncer are with Eaboratoire Ville Mobilité Transport, Ecole Nationale des Ponts et Chaussées, Institut Polytechnique de Paris, Uni Gustave Eiffel, Champs-sur-Marne, France; Mwendwa Kiko is with the University of Toronto, Canada.

Digital Object Identifier: 10.1109/IOTM.001.2400024

the energy supply, based on the requirements of the evaluations to be performed, i.e., in terms of evaluation criteria and description granularity. This step-based model enables inputs from the various stakeholders of the CAM ecosystem to be integrated in the design process. In addition, TRESSY embeds the specifics of a deployment context (i.e., a time horizon and a region) in the design of the scenarios by building on a definition of up-scale services tailored to CAM. This enables all aspects of a CAM service to be characterized, including the technical systems, the infrastructures, the business models and governance, in addition to the targeted users.

We elaborate on the specifics of TRESSY and illustrate how the approach can be used by applying it to the real use case of an evaluation study conducted as part of SAM, the French National Automated Vehicle pilot R&D project.[3] We demonstrate how stakeholders with various needs and requirements (researchers from different disciplines, automotive industry leaders, mobility service operators, local decision-makers) can jointly work on the design of scenarios for the systemic assessment of a system that builds on critical and converged infrastructures.

From Segmented to Holistic CAM Impacts Assessment

CAM is expected to have significant direct and indirect effects, ranging from micro vehicle-level, (e.g., on driving performance and safety) to macro networks and markets-level (e.g., on mobility demand behavior and supply efficiency) [8]. It is therefore essential to adopt a holistic approach for their assessment, which necessitates the implementation of a multi-factor evaluation framework and the analysis of mixed data sources.

Methods and Data Sources for CAM Evaluation

The effects of a CAM deployment translate into several direct and indirect environmental, public health, land use and socio-economic impacts. In a recent paper Tian *et al.* [6] identified a three-level classifications of CAM applications: vehicle-centric, infrastructure-centric and traveler-centric. This segmented approach to the impact assessment of CAM (Fig. 1) does not support the holistic evaluation of vehicle-infrastructure-service interactions as technical and service performance assessments use different methods and are often studied separately. Most of the studies focus either on the vehicle or on the service and tend to disregard the infrastructures [10].

CAM evaluations are also supported by different types of data sources. Figure 2 presents a nomenclature of typical data sources used for the evaluation of CAM. Evaluation data is either extracted from existing source (existing data) or produced as part of an evaluation study (new data). Existing data includes census data, previous travel survey data, open datasets. New data is either the result of a technical experimentation (CAM pilots, simulation) [3] or extracted from scenario-building studies. In a technical experimentation, the data is collected from the testing of automated vehicles (AV) or their components without any projection to another context (e.g., future time, different regions). In contrast, in a study based on scenario-building, evaluation data is the projection of one context to another. The process can be quantitative or qualitative, can apply creativity or codified knowledge, and can use numbers or narratives [4].

	Technical system evaluation frameworks	Service evaluation frameworks
Measure of Effectiveness	Vehicle performance Safety Traffic impacts Environmental impacts (air pollution, noise)	User behavior (human-machine interaction) User experience Accessibility Socioeconomic impacts Distributional impacts (social, spatial) Urban planning and design
Methods	Driving simulation Road test data analysis Traffic simulation Target crash population Life-cycle analysis	Technology Acceptance Model Cost-benefits analysis Qualitative and quantitative surveys Willingness to pay surveys Stated preference surveys Modelling

FIGURE 1. Overview of CAM applications and evaluation frameworks.

Scenario-Based Data Sources

Scenario-based data sources can in general be classified in two main families:

1. Technical **or** service
2. Technical **and** service

Technical and service data are used independently and not integrated in the former, while in the latter, two types of data are produced and consolidated through the study.

Technical and service detached: This configuration does not take into account how the evaluated service performance is achieved. This is the case in most highly cited publications that investigate AV through the lens of social science [11] and that define AV service characteristics without considering the supporting technical paradigm.

Technical and service attached: In this case, technical and service data results are consolidated based on two possible models. The first one is a technical to service data model. In the L3Pilot project[4] for instance, no original service is designed. The up-scaling of the impact of the technology is evaluated in private vehicles. The second is a service to technical data model. The emphasis is on the service that serves as constraint for the technical system: systems are matched to a specific region and specified based on the requirements of a concrete service offer. Technical performance is determined from the definition of service performance and matches the service objectives. In contrast to the first model, experiments cannot directly be used to extract results, and resorting to scenario design is necessary.

TRESSY: System Up-Scaling and Evaluation

Recent developments in CAM technology have been accompanied by public-road experiments[5] to collect vehicle, infrastructure and service performance data. These experiments are however still limited in fleet-size and spatial scale, and do not provide a sufficient basis for an upscaled evaluation of impacts. Scenario design for upscaled developments of CAM is as such required. In this section, we introduce TRESSY (Time, REgion, Service, SYstem), an approach to build up-scale scenarios of CAM services (considering the triptych vehicle-service-infrastructure) for multi-factor evaluation.

[3] https://librairie.ademe.fr/mobilite-et-transport/289-sam.html; accessed 05-08-2024.

[4] https://l3pilot.eu/; accessed 25-01-24.

[5] https://www.connectedautomateddriving.eu/test-sites/; accessed 25-01-24.

cenarios are useful to identify key trends, stakeholders and imperative requirements for the innovation to scale up, and the future challenges of upscaling.

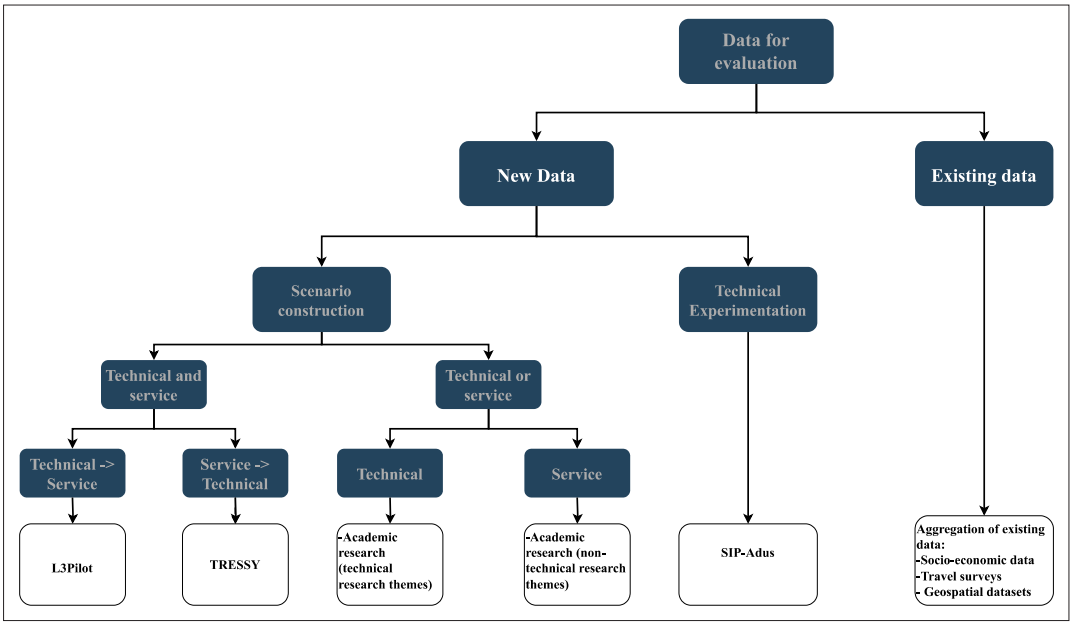


FIGURE 2. Data sources used in different studies evaluating CAM.

Scenario-Building and Up-Scaling Definition

Scenario-building analysis provides relevant tools to address complex and multidimensional issues in uncertain trends, and support various tasks (e.g., planning processes, transport development and land-use integration [5]). It can be used as a complementary forecasting tool for the future up-scaling of technological innovations. Scenarios are useful to identify key trends, stakeholders and imperative requirements for the innovation to scale up, and the future challenges of upscaling. Developing future scenarios requires a cross-section expertise and the involvement of multi-stakeholder to identify the key drivers of change and the uncertainties in the mid- or long-term [5].

Here we propose a new definition of up-scaling adapted to the context of CAM pilot projects that target real-world services. It includes four main aspects:

- The shift from a technology experiment to a mobility service that meets social needs;
- The spatial broadening of experiments;
- The strengthening of the network of actors around the technology;
- The integration within an existing mobility system.

We use this definition to design TRESSY.

TRESSY in Four Steps

TRESSY decomposes the design of an up-scale scenario in four steps:

1. Identification of a Time horizon for service deployment
2. Characterization of the REgion for service operation
3. Definition of the specifics of the operational Service
4. Description of the technical SYstems of the service

Scenarios are represented as a set of specifications that constitute input data for multi-factor evaluation studies. An overview of TRESSY is depicted in Fig. 3.

The objective of the first two steps is to contextualize each scenario both in time and space. These are critical choices as they condition the setting for service deployment and operation. The two steps ensure that the defined scenarios are in response to clearly identified mobility needs that the connected and automated technology contributes to address in order to be consistent with the definition of up-scaling, e.g., by reducing the operation costs of the service. The third and fourth steps focus on the features of the services, including how it is operated and financially supported, as well as the technical systems and infrastructures needed to support their delivery.

Step 1 — Temporal landscape for service deployment: TRESSY defines up-scale scenarios of CAM services in the context of a temporal horizon for (realistic) deployment and operation of the AV technology. The choice of a time horizon is a determining factor in the definition of a scenario. It factors in three main constraints. For the scenarios to be relevant, the temporal horizon needs to be aligned with the objective(s) of the evaluation studies. Evaluating the environmental impact of a small-scale field experiment (i.e., short term horizon) is likely to be insufficient to extract consolidated observations. In addition, the temporal horizon needs to be in line with the time frame of mobility plans to correspond to concrete transportation needs. Finally, the time horizon needs to be relevant to the evolution of the technical systems and their maturity in supporting operational services.

AVs are fast evolving technologies with possible disruptions in the course of their development. The farther into the future, the larger uncertainties, and the more scenarios are needed. Most of the scenario narratives on future urban mobility set medium-term (10–15 years) or long-term projections (20–30 years) [15]. Medium-term timeframes such as 2030 are reasonably challenging and makes it possible to reason upon existing trends (demographic, technology [5]. The risks of break-out in technological development can make up-scaling scenarios of AV services obsolete in longer term horizons.

Step 2 — Region characteristics for service operation: This step aims to specify both the environment of service development (i.e., urban, suburban, rural) and its dynamics at deployment time (e.g., socio-demographics, travel patterns). Three main aspects are needed to characterize the region. The first concerns the evolution of the population demographics and the environment (type of roads, topology) within the defined time horizon. While the former determines mobility patterns, the later conditions the type of services that can be offered. In particular it is crucial to identify planned development work within the region and review their consequences for transportation services. The second aspect is to understand the demand for transportation at the projected horizon. It involves reviewing existing mobility options in order to determine whether future services need to complement or replace existing supply, and their anticipated deployment scale. The third aspect is to understand the ecosystem and supporting business models by identifying key stakeholders in the region, including decision-makers, industrial actors, operators of transport, service providers, the public, etc., so as to extract insights on the cost structures for the implementation of these services.

Step 3 — Design of the mobility service: The characteristics of the region identified in step 2 serve as drivers for designing realistic services in the third step. This is achieved in two phases. The first phase, referred to as *exploratory*, focuses on developing a set of business models of mobility service scenarios. It uses the Business Model Canvas [13] to identify key activities and resources, customer segments and relationships, as well as the operational model needed to deploy such scenarios. The second phase, referred to as *definition*, selects from the obtained set of scenarios the ones with the highest potential in terms of

1. Satisfying the identified needs in step 2
2. Being realistically cost-effective for deployment

The set of selected scenarios are used to specify in detail the characteristics for the associated services.

TRESSY defines each service based on a taxonomy of service attributes as depicted in Fig. 4. Attributes are organized in three general classes and twelve subcategories. The usage class encompasses all attributes that pertain to the service demand, pricing policy, accessibility and quality. It specifies whether the service is public, if it is collective, the profile and number of passengers that are anticipated, the rates, if it comes with a booking option etc. The supply class includes all attributes associated with the geographical coverage (e.g., type of stops, distance between stations) and timetable of the service, as well as the attributes associated with the integration of the service to other networks (i.e., inter-modality) and with the infrastructures needed to support its operation (e.g., location of the depot, type of lanes needed). The fleet class spans all attributes that are used to define the service in terms of its fleet size, the type of vehicles that it employs, the system automation and their performance characteristics.

Step 4 — Description of the technical systems: The last step is concerned with describing the technical characteristics of the infrastructures needed to support the services. Technical specifications in TRESSY are decomposed in five components, namely vehicle, supervision, autonomous driving (AD) provider, data flows and infrastructure, as shown in Fig. 5.

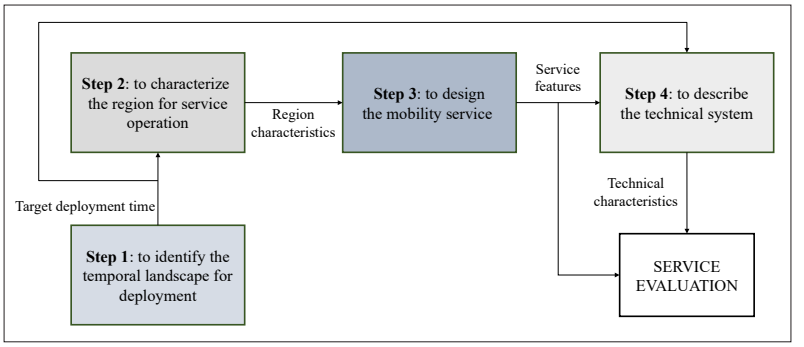


FIGURE 3. Overview of TRESSY.

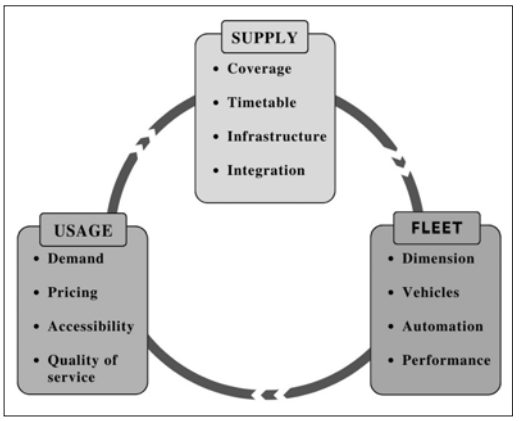


FIGURE 4. Classes of service characteristics.

The vehicle component covers the vehicle operations (e.g., specifications of the engine, battery, tires) as well as the automated system (i.e., characteristics of the on-board sensors and computers) and its performance (i.e., in terms of perception and resulting safety margins). The supervision component includes information about both the human operators responsible for monitoring, maintenance and reactive intervention tasks, as well as the supporting systems such as cameras and dedicated computing services. The AD provider component is specific to the characteristics of the compute resources needed to assist automated driving in terms of tracking lists, vehicle logs, on-board software, etc. The infrastructure component concerns the digital environment of the mobility service, including communication equipment (antennas, road side units), resources for perception (cameras, lidars), infrastructure for data management (edge *vs.* centralized datacenters) and road facilities for controlling traffic (through connected traffic lights or connected gates for instance). In addition, TRESSY specifies the main types of data exchanges (communicating entities and volume of traffic) that need to be orchestrated to support automated driving and service operations.

Characteristics under each component are defined based on a set of quantitative parameters, e.g., a computing service is characterized by the number and type of servers, that can be expressed as a range of values and their evolution at the targeted time horizon (i.e., major technological developments, mature technology). While TRESSY fixes these characteristics as reference for

The passive road infrastructure is upgraded for automated driving with doubled road surface marking and increased frequency (x1.5) for on and off road maintenance and cleaning interventions.

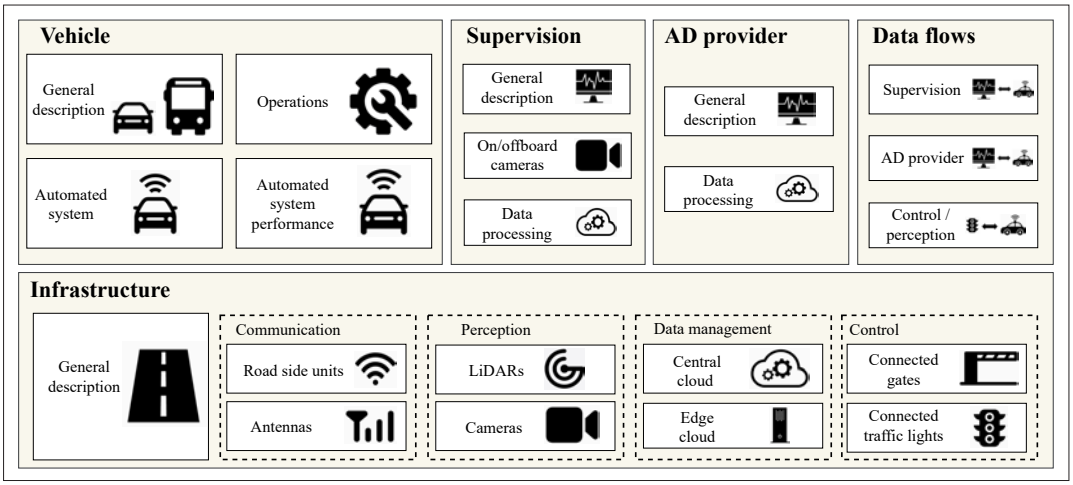


FIGURE 5. Technical system components.

a model of an up-scale scenario, the granularity of the description (i.e., set of parameters) needs to be adjusted to the requirements of the evaluation studies. Each technical system is described in a modular fashion, which enables the representation of different types of up-scale scenarios, e.g., from operating a fleet of autonomous vehicles with no intelligent infrastructure to automated vehicles heavily using augmented perception.

Use Case Illustration

We applied TRESSY in the context of the pilot project SAM.[6] The objective of the SAM project is to explore different types of technical systems (vehicles and infrastructures) for CAM by running AV experiments in several French cities with a focus on public transport services, private mobility and automated urban logistics. The project involved multiple stakeholders, including AV manufacturers, transport operators, infrastructure managers, suppliers and academic researchers. Services were evaluated in terms of different factors, including acceptability, safety, user behavior, socioeconomic and environmental impact, both on experimental setups and on designed up-scale scenarios. We used TRESSY to construct these up-scale scenarios.

Scenarios Overview

Two use case scenarios for the deployment of CAM services at a 2030 time horizon were developed for the Paris-Saclay region, a suburban area of 187 km^2 located in the South area of Paris. The region hosts residential neighbourhoods, company and university campuses, as well as shopping areas. The 2030 time horizon was chosen to be aligned with the French National strategy for the deployment of AV and also gained the consensus of the industrial partners of the project with respect to technological maturity for realistic deployments. Each scenario consists in a combination of a service and a technical system as per Step 3 and Step 4 of TRESSY.

Scenario 1: Represents an on-demand, point-to-point service operated with a fleet of 200 light weight vehicles, each with a capacity of 6 to 9 seats and a space for a wheelchair. The service is designed to be complementary to the existing public transport offer (regular buses and train lines in the area). Passengers are picked up from meeting points distributed every 300 to 500 meters on the covered area. The service is integrated as an option within the Mobility-as-a-Service (MaaS) platform of the Paris-Saclay region.

Scenario 2: Consists in a 15 km fixed line minibus service that operates in a hybrid mode with fixed time passenger pickups at existing bus stops during off-peak times and on-demand reservation from the same stations at night. The service is complementary to an existing bus service.

An overview of the service coverage for the two scenarios is depicted in Fig. 6. The blue line on the map represents the service of Scenario 2, while the service of Scenario 1 is available over the area bordered in red. Due to space limitation, we only elaborate on the technical system of Scenario 1 in the next subsection. Scenario 1 is more challenging to deploy than Scenario 2 as it covers a larger region with multiple possible routes.

Technical System: Case of Scenario 1

As shown in Fig. 5, the technical system associated with the deployment of a CAM service involves multiple components. We present here the specifications developed for the service of Scenario 1. The detailed description of all system parameters is available as a technical report in [7].

Vehicle specifications: The vehicle operates at level 4 of automation [14] with no safety driver on-board. Vehicle control is event-based only (e.g., in case of an incident) and triggered remotely through a supervision centre where an operator can take over. The vehicle is equipped with a set of on-board sensors and cameras, one on-board unit to record driving data, two on-board computers and a human-machine interface (a screen for instance) for passenger information and security calls. The vehicle is full electric, with one engine and a 50kWh battery. This enables a 10 hour autonomy (or equivalent driving range of 200 km). Charging is done at a depot where a bay of 100 parking spaces with 22kW charger is provisioned. Upon charging and maintenance, the vehicle downloads software upgrades, with an estimated volume of data of 1GB a month, and uploads to the supervision system of the service operator, service activity logs that record unanticipated stops, incidents, route mapping errors, etc., with an estimated volume of data of 1GB a day.

[6] See earlier section.

Road infrastructure specifications: The vehicle operates in mixed traffic conditions with a maximum speed of 110km/h. The possible routes that the vehicle can take are pre-trained over 2,000 km and recorded in a high definition map that is updated on a daily basis based on the received service activity logs (estimated 100MB of data downloaded a day). The passive road infrastructure is upgraded for automated driving with doubled road surface marking and increased frequency (x1.5) for on and off road maintenance and cleaning interventions. In addition the road is augmented with a connected infrastructure involving 200 connected traffic lights (with 4 connected traffic lights per road intersection), 210 road side units (approximately one every 10km), road side cameras (approximately 4 every 10km) and road side lidars to guide driving in case of low visibility and/or when pedestrians or two-wheels are detected in close distance. The connected infrastructure enables vehicle-to-infrastructure (V2I) and infrastructure-to-vehicle (I2V) communications. V2I is used to report information collected by the vehicle about accidents, road closure and road work, while I2V is used to secure intersection crossing. Vehicle-to-vehicle (V2V) communication is also enabled through CV2X 5G to signal upstream incidents to other passing by V2V-enabled vehicles.

Communication infrastructure specifications: A communication infrastructure is required to support teleoperation, as well as for localization. Teleoperation is achieved through relaying via 5G antennas deployed in the covered area. Localization is performed using global navigation satellite system (GNSS) and error correction via real time kinematic (RTK) information, and necessitates the deployment of GNSS repeater,s especially in the forested zones of the service and long tunnels (longer than 100m).

Supervision specifications: The supervision enables remote control of the vehicles through a supervision centre where operators are trained to oversee a CAM service. An operator can control up to five vehicles. In addition eight vehicles are reserved for immediate interventions in case an incident occurs in the service. Information is pushed to the vehicles operating the service, regularly for passenger information and on a per-event basis for instance to instruct about alternative routes if an incident is detected. Each vehicle also sends information to the supervision system, including monitoring data about the vehicle performance and real-time data of the status of the service that is used to be displayed as service information on the MaaS platform.

Discussion

TRESSY worked well for identifying the needs and challenges of CAM services aligned with the experiments of the project. Several observations could be made from the application of the proposed method.

A practical tool for workshop activities: The two scenarios were developed as part of six workshop activities. Each workshop gathered around 15 participants from the project, including representatives from AV manufacturers, transport operators, infrastructure managers, local public authorities and academic researchers. To maximize the collaboration between the stakeholders, participants were provided with a spatial diagnosis of the region and a set of personas, i.e., fictional inhabitants and potential users of mobility services in 2030 [5], at the beginning of each workshop. These inputs were essential to contextualize the services and their technical systems. Participants designed the general architecture for each scenario by combining their expertise based on the service requirements (Fig. 4) and the pre-identified technical components (Fig. 5). The resulting detailed technical system descriptions constituted a common knowledge for industrial and academic partners to share for further evaluations.

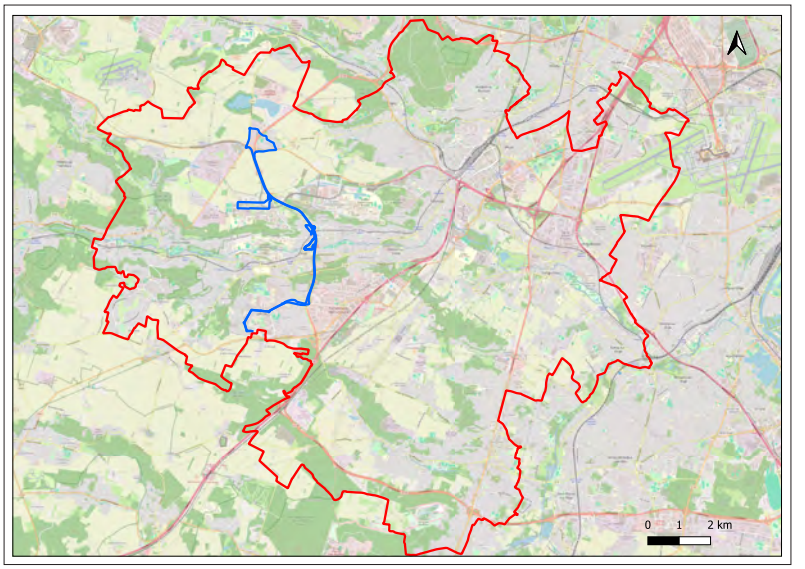


FIGURE 6. Envisioned service deployment area.

Importance of local authorities participation: The workshop participants had to identify the future contribution of CAM services compared to services with drivers. Discussion between experts from different fields was useful to pinpoint potential blocking points for the development of CAM services, and the conditions to overcome them. The participation of the local authorities enabled the alignment of technological development with the needs of a fully functional service anchored in a specific region. Inertia in critical infrastructure development could be visualized when local strategies were in contradiction with technological needs. As technology developments and local deployments can follow different timelines, bringing together local authorities and technological stakeholders is determining to reach a consensus.

Data challenges: The collection of data raised three main concerns. The first relates to trade secret. In a multi-stakeholder project, access to data can be an issue given that industrial stakeholders are often reluctant to share information about their technologies and services. We observed that TRESSY facilitated the resolution of this issue by targeting a consensus on the future data that can be used for evaluation without compromising trade secret. Stakeholders were more keen on giving insights on future data during the discussions at the workshops when other participants also shared their perspectives. The second issue was data quality, which needed to satisfy the requirements of the evaluation in terms of completeness, accuracy or/and granularity. During

our workshops, the participants were asked to assess the expected monetary and operational values of the developed CAM services. While we collected ranges of values for each parameter, the orders of magnitude were not systematically associated with the right supporting argument. Some participants also did not feel qualified enough to comment on the data, which shows that selecting the relevant people from the ecosystem is critical to interface scenario-design and evaluation. The third issue concerns post-workshop data processing, to determine missed value or correct incoherent results. In the project, the final decision was taken by the organizers when no consensus emerged or when no value was provided, and was based on the literature.

Conclusions

The paper introduces TRESSY, a novel scenario-building approach to explore potential up-scaling for the deployment of CAM services and their associated IoT-enabled and critical infrastructures. Innovative projects are often experimented at small scales, where user behaviors and interactions with other systems and existing infrastructures are not representative of future large-scale deployments. TRESSY facilitates the development of diversified and collaborative scenarios of up-scaled CAM services that can be used for multi-factor evaluation. The proposed approach enables inter-stakeholders and inter-disciplinary work on the development of connected and automated vehicle technology, especially in the context of experiment-based projects where the expectations of public and private stakeholders can differ. The approach helps a heterogeneous group of stakeholders engage in determining points of divergence and reach a consensus in the type of services and infrastructure setups needed to support CAM. We are currently working on the implementation of an online tool to support the organization of TRESSY-based workshops remotely and plan to re-run the activities with a greater diversity of stakeholders, including members of the public, that will enable us to assess the barriers to an application of the approach to larger and more sophisticated CAM implementations.

References

[1] B. Hurwitz, and R. Charon, "A Narrative Future for Health Care," *Lancet*, vol. 381, no. 9881, 2013, pp. 1886.

[2] A. B. Moniz, "Scenario-Building Methods as A Tool for Policy Analysis," *Innovative Comparative Methods for Policy Analysis*, 2006, pp. 185–209.

[3] D. J. Fagnant and K. M. Kockelman, "The Travel and Environmental Implications of Shared Autonomous Vehicles, Using Agent-Based Model Scenarios," *Transportation Research Part C Emerging Technologies*, vol. 40, 2014, pp. 1–13.

[4] R. Ramirez and C. Selin, "Plausibility and Probability in Scenario planning," *Foresight*, vol. 16, no. 1, 2014, pp. 54–74.

[5] F. Vallet *et al.*, "Tangible futures: Combining Scenario Thinking and Personas — A Pilot Study on Urban Mobility," *Futures*, vol. 117, 2020, pp. 102513.

[6] D. Tian *et al.*, "Performance Measurement Evaluation Framework and Co-Benefit/Tradeoff Analysis for Connected and Automated Vehicles (CAV) Applications: A Survey," *IEEE Intelligent Transportation Systems Mag.*, vol. 10, no. 3, 2018, pp. 110–22.

[7] N. Kotelnikova-Weiler and A. Feraille Fresnet, "Analyse de Cycle de vie Des Systémes Techniques de la Mobilité Automatisée," *Technical Report, hal-03827916, Laboratoire Ville Mobili{é Transport*, Ecole des Ponts, Universit{é} Gustave Eiffel, 2022; https://hal.science/hal-03827916.

[8] D. J. Fagnant and K. Kockelman, "Preparing A Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations," *Transportation Research part A: Policy and Practices*, vol. 77, 2015, pp. 167–81.

[9] S. Sohrabi *et al.*, "Quantifying the Automated Vehicle Safety Performance: A Scoping Review of the literature, Evaluation of Methods, and Directions for Future Research," *Accident Analysis & Prevention*, vol. 152, 106003.

[10] F. Carreyrev *et al.*, "Economic Evaluation of Autonomous Passenger Transportation Services: A Systematic Review and Meta-Analysis of Simulation Studies," *Revue d'économie industrielle*, 2023, pp. 89–138, no. 178–79.

[11] L. Mora, X. Wu and A. Panori, "Mind the Gap: Developments in Autonomous Driving Research and the Sustainability Challenge," *J. Cleaner Production*, vol. 275, 2020, pp. 124087.

[12] B. Metz *et al.*, "L3Pilot Driving Automation: Deliverable D3.3 Evaluation Methods," Technical Report, 2019; https://l3pilot.eu/fileadmin/user_upload/L3Pilot-SP3-D3.3_Evaluation_Methods-v1.0_DRAFT_for_website.pdf.

[13] A. Osterwalder, and Y. Pigneur, *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers (Vol. 1)*, John Wiley & Sons, 2010.

[14] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2018; https://www.sae.org/standards/content/j3016_201806/.

[15] G. Lyons, and C. Davidson, "Guidance for Transport Planning and Policymaking in the Face of an Uncertain Future," *Transportation Research Part A: Policy and Practice*, vol. 88, no. 26, 2016, pp. 104–16.

Biographies

Manon Eskenazi (manon.eskenazi@enpc.fr) is a researcher in urban planning at Ecole nationale des ponts et chaussees (France). Her research focuses on mobility sociotechnical transitions (i.e., active mobility, digital apps and automated vehicles) through the threefold lens of policies, practices and space.

Fabien Kaptue Bopda (fabien.kaptue-bopda@enpc.fr) is a research assistant in mobility at Ecole nationale des ponts et chaussees (France). His interests focus on the role of digitalisation in the development of novel platforms for mobility.

Mwendwa Kiko (mwendwa.kiko@mail.utoronto.ca) is a third year Ph.D. student at the University of Toronto. His main research interest is modelling of the land-use – transport interaction and his thesis topic is the modelling of vehicle transactions for an integrated transport and land use model.

Natalia Kotelnikova-Weiler (natalia.kotelnikova@enpc.fr) is a former member of Laboratoire Ville Mobilité Transport (France), currently working as an independent researcher and consultant. Her work evolves around passenger transportation and its environmental impacts mainly through the evaluation of pollutants emissions, energy consumption and life cycle analysis at regional, urban and mobility service levels of conventional and innovative (automated) transport modes.

Daphne Tuncer (daphne.tuncer@enpc.fr) leads programs of research at Ecole nationale des ponts et chaussees (France). Her work focuses on the intersection between computer networks, information systems and data science for the management of converged cyber-physical system infrastructures (e.g., in support of mobility and energy). She is the recipient of a 2018 Imperial College Research Fellowship and the 2021 IEEE CNOM Young Professional award.

IEEE Collabratec™

Bright Minds. Bright Ideas.

Introducing IEEE Collabratec™

The premier networking and collaboration site for technology professionals around the world.

IEEE Collabratec is a new, integrated online community where IEEE members, researchers, authors, and technology professionals with similar fields of interest can **network** and **collaborate**, as well as **create** and manage content.

Featuring a suite of powerful online networking and collaboration tools, IEEE Collabratec allows you to connect according to geographic location, technical interests, or career pursuits.

You can also create and share a professional identity that showcases key accomplishments and participate in groups focused around mutual interests, actively learning from and contributing to knowledgeable communities. All in one place!

Network.
Collaborate.
Create.

Learn about IEEE Collabratec at
ieeecollabratec.org

IEEE

Enhancing Resilience in IoT Water Systems Using Data-Intelligence and Decentralization

Haitham Mahmoud, Wenyan Wu, Mohamed Medhat Gaber, and Yonghao Wang

Abstract

In recent years, concerns regarding the security of water networks have escalated due to the increasing integration of water assets (actuators and sensors) with the Internet, combining Information Technology (IT) and Operation Technology (OT). This integration promises improved services for water networks but also introduces the risk of cyber-attacks and physical threats. As a result, there is a growing need for novel security measures to protect integrated Cyber-Physical Systems (CPS) in water distribution systems (WDSs). This article assesses actual incidents and potential Cyber-Physical (CP) attacks on water systems, explores their operational impacts, and suggests mitigating measures. It introduces a secure architecture for an integrated CPS in WDS. The study incorporates attack detection and data validation models to enhance system robustness and reduce risks, adhering to the security criteria of Water 4.0. First, the attack detection model utilizes a two-stage architecture employing six Machine-Learning (ML) algorithms, resulting in developing a simulation model with the best-suited configuration. Second, the data validation model uses blockchain technology on transmitted data, creating a simulation model for water consumption data with various input types, consensus mechanisms, and data output conversion methods. Finally, this article provides a foundation for researchers, professionals, and operators in the water sector to experiment with, evaluate, and further develop this secure architecture for their water systems. Simulating their networks using the proposed architecture allows them to identify the most suitable configurations and parameters for their specific implementations.

Introduction

The rapid development of smart water networks, also known as Water 4.0, which seamlessly integrate IT and OT, has created an immediate demand for improved security in WDS. While this integration can potentially optimize water infrastructure and services, there is also an increased risk to safety. Water supply, water treatment, water distribution, and water sanitary removal (wastewater) are the four vital components of the water supply and distribution systems. The water supply system called the supply-side water distribution system, provides households with treated water transported from various sources, including reservoirs, dams, aquifers, wells, and aqueducts. The main goal of water treatment is to remove biological contaminants using filtration processes. The water distribution system uses smart meters, water tanks, and pumps to make transporting water through pipelines easier. Finally, sewers and subsystems are used in the sanitary and wastewater removal system to move untreated water to treatment facilities.

Since many of these systems date back to the late 1800s, they have aged and may require replacement or repair, potentially making them susceptible to disruptions and security threats. Protecting modern smart water networks necessitates controlling public access and implementing advanced security measures. In the context of the Smart Water Network (SWN), consisting of five layers - physical, sensing and control, collection and transmission, data management and presentation, and data fusion and analysis, cyber-physical systems are integrated into water infrastructure, enhancing automation, data analysis, and safeguarding against both cyber and physical threats. The risk is further highlighted by a 2019 RiskIQ report, revealing that cybercrime costs over $1.1 million every minute and affects more than 1,800 individuals by causing infrastructure damage and service disruptions [1]. The ICS security market, exhibiting a Compound Annual Growth Rate (CAGR) of 7.6% from 2014 to 2019 and witnessing an increase in investments for infrastructure and asset protection from $7.82 billion to $11.29 billion, underscores these vulnerabilities [2].

A secure smart water management system is needed to defend against ever-evolving cyber threats in contemporary water distribution networks. However, studies continue to focus on attack detection, accuracy and false positives due to a lack of existing simulators. Creating a method that minimizes false positives in an integrated threat detection system is essential to improving infrastructure security. Moreover, reliable data verification is necessary for smart water management. The literature on this subject lacks in-depth implementation. Moreover, a comprehensive investigation into data verification is lacking, which exposes a security flaw. Implementing an effective data verification system is essential to provide an additional line of defense against potential attacks.

A resilient smart water management system is proposed to detect and localize intrusions and

Digital Object Identifier: 10.1109/IOTM.001.2300272

The authors are with Birmingham City University, UK.

enhance infrastructure resilience. Data intelligence and decentralization are key components of the system. This system uses blockchain technology for secure measurement verification inside sensor devices and ML for attack detection and localization.

- An attack detection system is proposed that utilizes six state-of-the-art ML algorithms: Isolation Forest (IF), Bag of SFA-SAX Symbols (Boss), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost). Moreover, four different-size WDS are included in the evaluation: Net1, Mini-town, C-town, and D-town. This assessment examines characteristics including accuracy, detection time, and time-to-detect while thoroughly testing the system's functionality under four attack types.
- The data verification system is proposed using blockchain technology. Two alternatives are investigated for data entry:
  1. Time series data relevance
  2. Integration with one timestamp using a water modeling called EPANET.

  Moreover, four different consensus mechanisms working within the same water distribution systems are included in a thorough analysis. The system is evaluated based on the number of transactions, generation and mining time, and throughput, systematically assessing the performance of each mechanism.

This article is organized as follows: we present the threat model and the existing study. We propose the architecture by discussing the key components, attack detection and visualization, simulation and GUI in the system. We discuss the results and the future research directions. We aim to address the future research directions. Finally, we conclude the work.

Background

Threat Model

We aim to implement a secure smart water network with two primary objectives:

1. To mitigate CPS attacks on WDS
2. To validate shared data on the WDS within decentralized connected networks

We have developed a threat model that considers honest network peers to achieve these objectives. Effectively addressing potential threats and vulnerabilities is crucial for securing smart water networks. Robust access controls and authentication are essential to prevent unauthorized data access and safeguard sensitive information. Data integrity checks and encryption are critical as data tampering can compromise information integrity. Intrusion detection and device authentication are necessary to counter the infiltration of malicious nodes, including compromised Internet of Things devices. Real-time monitoring and intrusion detection systems are indispensable to avoid prolonged breaches due to delayed attack detection. Redundancy and proactive maintenance are vital to minimize the impact of service interruptions on customers. Considering centralization risks and ensuring transaction transparency are also important factors. In addition, verifying the data within a smart water network is crucial to ensuring its reliability and accuracy.

	System	Purpose	Tech. used			Impact		
			Blockchain	Secu. IoT	Traffic Analysis	Security	Privacy	Resilience
[7, 8]	Watering System	Watering Managment	√	√	×	×	√	√
[5]	Watering System	Encrypted Comm.	×	×	×	√	×	√
[6]	WDS	Secure WDS	√	√	√	√	√	√
[9, 10]	WDS	Attack Detection	×	×	√	√	×	√
[11, 12]	WDS	Data verification	√	×	×	√	×	√
This paper	WDS	Data verification and Attack Detection	√	√	√	√	√	√

TABLE 1. Evaluation of the existing studies that used technologies (i.e., blockchain, secure IoT or traffic analysis) for security, privacy or resilience impact.

Related Works

This section provides an overview of studies in water systems employing blockchain, ML, or IoT security to leverage security (as shown in Table 1). Blockchain has gained significant attention in watering systems and agriculture due to its ability to establish transparent, immutable water usage records. This ensures equitable water rights allocation and elevates the overall system security by safeguarding against data manipulation and unauthorized access. In parallel, other water systems use blockchain to promote water rights and transparency.

ML takes the lead in smart water systems, primarily dedicated to attack detection. These smart water systems encompass diverse domains, including distribution, supply, wastewater, drinking water, and irrigation. Some studies use ML and blockchain for multifaceted applications like seed monitoring, leakage detection, and predictive maintenance [3]. However, it is crucial to highlight that the paramount focus in these integrated systems remains on enhancing security, with particular emphasis on identifying and thwarting potential attacks, thus creating resilient water infrastructure. Furthermore, blockchain and ML are important water distribution systems and irrigation technologies. Some studies have ventured into encryption for securing data transmission and enhancing data privacy and integrity [5, 6]. However, it is noteworthy that the primary objective of these investigations centers on leveraging the security and enhancing the resilience of water infrastructure, with relatively limited attention given to addressing privacy concerns.

This highlights the insufficiency of existing studies that comprehensively leverage the security potential presented by the convergence of IoT security, ML, and Blockchain.

Proposed Architecture

The proposed WDS architecture comprises four core components: the WDS model on EPANET, data validation system (using blockchain technolo-

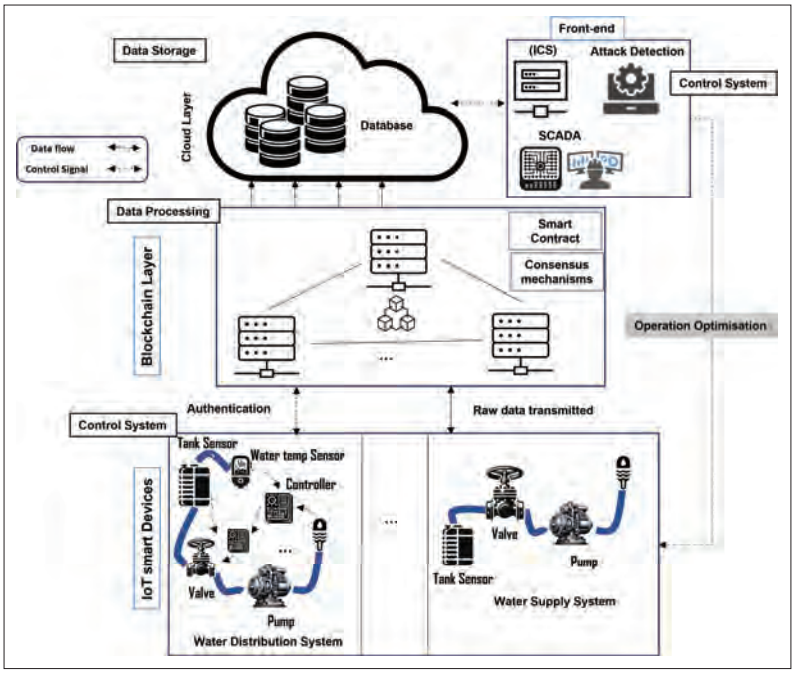


FIGURE 1. The suggested architecture of secure IoT-WDS to enhance water resilience using Data intelligence and Decentralization. The architecture focuses on two key components which are data validation using blockchain and attack detection using ML.

gy), cloud/storage, and system services (as shown in Fig. 1). The EPANET WDS model allows users to design water distribution systems and is known for its adaptability and integration capabilities. The data validation process validates transmitted measurements from WDS assets like water tanks, using various methods, forming a chain of hashed data. Metadata is stored in the cloud for reference, and both data chains and metadata are compared using a fundamental hash function. Finally, data analysis and processing occur through system services applications, including SCADA, ICS, and attack detection.

Key Components

Use-Cases: Four use cases cover a variety of WDS configurations and characteristics. With a granularity of one hour, Net1 comprises 11 nodes, 12 pipelines, one pump, one tank, and one reservoir. Its duration is 336 hours. Mini-town has ten nodes, ten pipelines, two pumps, one tank, and one reservoir. It runs for 168 hours at a granularity of 20 minutes. C-Town is a 72-hour operation with a 6-minute granularity that consists of 429 pipelines, 11 pumps, seven tanks, one reservoir, and 396 nodes. Finally, D-Town uses 407 nodes, 443 pipelines, 11 pumps, seven tanks, and one reservoir over 72 hours, albeit with a 20-minute granularity.

IoT Devices and Authentication: Cryptographic techniques are used for authentication between blockchain networks and IoT devices. Every IoT device is equipped with an individual public-private key pair (*DeviceKey_public*, *DeviceKey_private*), in which the private key is safely stored (*DeviceKey_private*). The blockchain network creates a cryptographic challenge (*Challenge*), essentially a hashed value of particular data, and sends it to the device in response to an authentication request. The machine uses its private key to answer the challenge with a response (Response): Response = Hash(*Challenge*, *DeviceKey_private*). The blockchain network verifies the response by hashing the challenge using the public key linked to the device's identity and ensuring that Response = Hash(*Challenge*, *DeviceKey_public*). When verification is successful, the device is verified, access is granted via smart contracts, and an immutable audit trail is created for security and transparency.

Encryption and Data Signature: A combination of encryption, hashing functions, and digital signatures is used to ensure messages are secure during transmission. The sender encrypts consumption data using the recipient's public key, ensuring that only the intended recipient can decrypt it with the private key. The blockchain generates digital signatures that can be used to verify the authenticity and source of data. Altering the data changes the digital signature, preventing potential attackers from manipulating consumption amounts. A bloom filter and zero-knowledge proof verify the blockchain's integrity. Zero-knowledge proofs verify data consumption without revealing customer identity, while bloom filters match pseudonyms with block data. Blockchain security and trustworthiness are ensured through the verification process overseen by a mining node.

Data Validation and Blockchain: These are pivotal elements of any blockchain system, overseeing communication and verification processes through consensus mechanisms. This study introduces five consensus mechanisms — Proof of Work (PoW), Proof of Trust (PoT), Proof of Assignment (PoA), and Proof of Vote (PoV) to accommodate fast-action, voting-based, and real-time verification in the context of water distribution systems. PoA is recommended due to the WDS phenomena' fast-paced and voting-based nature. PoW, the initial consensus mechanism derived from Bitcoin, requires all network peers to validate transactions before rewarding the first verifier. PoT relies on the peer with the highest reputation for data verification, employing a trust matrix with reputation increments for approved data. PoV uses a voting system for data verification, requiring over 3/4 of the feedback to indicate genuine data. PoA offers low-processing and fast verification, suitable for IoT systems with less stringent security needs.

Cloud Layer and Front-End APIs: When storing data on a blockchain, it is usually not saved directly but rather as a hash to the original data. A cloud-based storage system occupies the actual data. The hash value is a link to the cloud data and is used for verification. Access to the cloud, where the data is stored, is necessary to recover the original data. The hash value and the reference to the original data are all on the blockchain; the actual data is not there. When data is accessed on the blockchain, the data is retrieved and matched from the cloud using the recorded information on the blockchain.

Establishing seamless connectivity with front-end applications, such as SCADA and ICS, is essential for improving the resilience of intelligent water management. One of the primary concerns of this article is the attack detection system. By achieving seamless integration and prioritizing attack detection, we aim to enhance the system's ability to monitor, control, and optimize water-related processes, ensuring the security and reliability of our critical water infrastructure.

Attack Detection and Visualization: This study comprehensively evaluates attack detection systems and their resilience against malicious intrusions. Due to the prevalence of sensor tampering in critical infrastructure sectors like WDS, this article explores four different WDS scenarios and generates and assesses attacks against them. The effectiveness of SVM, KNN, RF, Boss, XGBoost, and IF algorithms in countering these attacks is assessed. The SVM specializes in identifying patterns, assisting in the detection of abnormalities that may indicate a possible attack. KNN assesses data point proximity, which aids in the discovery of anomalies in water system behavior. RF improves attack detection by combining multiple decision trees with its ensemble learning technique. Boss, a symbolic representation-based program, detects and analyzes complicated patterns in water system data, revealing possible risks. As a strong gradient-boosting algorithm, XGBoost iteratively refines its understanding of system dynamics, resulting in better attack detection. Finally, IF's ability to separate anomalies is useful in identifying odd occurrences or attacks by differentiating them from the majority of typical system behavior. Together, these classifiers offer a robust protecting water system from potential security threats.

Assessment time, accuracy, precision, and F1 score are evaluation metrics used in the performance assessment of attack detection systems. Assessment time measures the time necessary to identify potential attacks in water systems. Accuracy indicates the system's overall correctness in detecting both attacks and non-attacks. Precision quantifies the system's ability to identify attacks while minimizing false alarms correctly. The F1 score provides a balanced assessment of precision and recall, considering both false positives and false negatives in attack detection.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Simulation and GUI

Data entry in the tool operates in two distinct modes: direct integration with EPANET, a prominent water modeling software, and the importation of time series data from CSV files. The direct integration is designed for real-time data processing, involving the transmission of individual timestamps. This differs from the importation of time series data, which handles pre-extracted time series data from EPANET, encompassing a wealth of datasets for comprehensive analysis.

Time series data is used, either directly from CSV files in MATLAB. The attack detection API extensively studies two detection techniques — self-supervised and unsupervised. Users select their preferred method during parameter selection in the model's graphical user interface (GUI). Data training and testing in different training and testing approaches — each with unique benefits — is determined by the selected method. Once time-series water data has been analyzed, attack origins can be identified via an attack localization technique. Users give information about the consensus method and data source (water distribution model or Excel file). Hashing-based validation (such as PoW, PoA, PoT, and PoV). Recognition of the chain or rejection for security-related reasons is established by data authorization. While the data validation model shows chained and dropped data and evaluates consensus process performance using multiple coefficients (e.g., latency, number of blocks per transaction), the attack detection model visualizes attacked assets and evaluates performance across data output coefficients (e.g., accuracy, precision, time-to-detect).

An intuitive GUI is developed using MATLAB, featuring user-friendly drop-down menus for dataset selection and user-activated buttons for ML algorithm selection. Moreover, the water distribution systems are visualized, emphasizing the precise localization of the attacks through highlighted sections.

WDS	#	Detection Algorithms					
		RF	SVM	KNN	Boss	XGboost	IF
Net1	1	🛡	🛡	🛡	🛡	🛡	🛡
Mini-town	2	🛡	🛡	🛡	🤖	🛡	🛡
C-Town	2	🛡	🤖	🤖	🛡	🛡	🛡
	3	🛡	🤖	🤖	🤖	🛡	🛡
	4	🛡	🤖	🤖	🤖	🛡	🛡
D-Town	2	🛡	🤖	🤖	🤖	🛡	🛡
	3	🛡	🤖	🤖	🤖	🛡	🛡
	4	🛡	🤖	🤖	🤖	🛡	🛡

TABLE 2. Evaluation of Attacks Detection towards four water modelling systems (Net1, MIni-town, C-Town, and D-Town), in which 🤖: Failed to detect and 🛡: detected the attack.

Discussion

Security Resilience and Attack Detection

Using EPANETCPA in MATLAB, data manipulation attacks are created within water distribution systems [13]. Four attacks are generated for security assessment as the following

- #Attks1 and 2 involves manipulate water tank sensor (T7) control signal before sending it to the SCADA. Through signal alteration, a potentially drought-causing false HIGH signal is falsely indicated when it is LOW.
- #Attks3 and 4 display inaccurate water level information, leading to operational failures like leaving the pump running and creating flooding.

Figure 2 and Table 2 present the assessment results, including evaluating assessment time, accuracy, precision, and F1 score across six algorithms for four datasets. With increasing dataset size, XGBoost, Boss, and SVM show the longest assessment times. Boss, XGBoost, RF and IF algorithm reaches 5.7, 5.2, 4.9 and 4.5 seconds for the C-Town Dataset, indicating higher computational costs associated with more complex models. Regarding accuracy, all algorithms achieve 100% or 0% based on detected attacks, with Net1, Minitown and D-Town datasets showing nearly 100% accuracy, except when using SVM for Minitown, while D-town exhibits lower accuracy. Boss performs 35% accuracy for the C-Town because of its limitation in detecting attacks in the C-Town dataset. SVM performs the least effectively in preci-

Implementing blockchain at the water tank level safeguards sensing and control signals, preventing unauthorized alterations that could lead to potential flooding or drought.

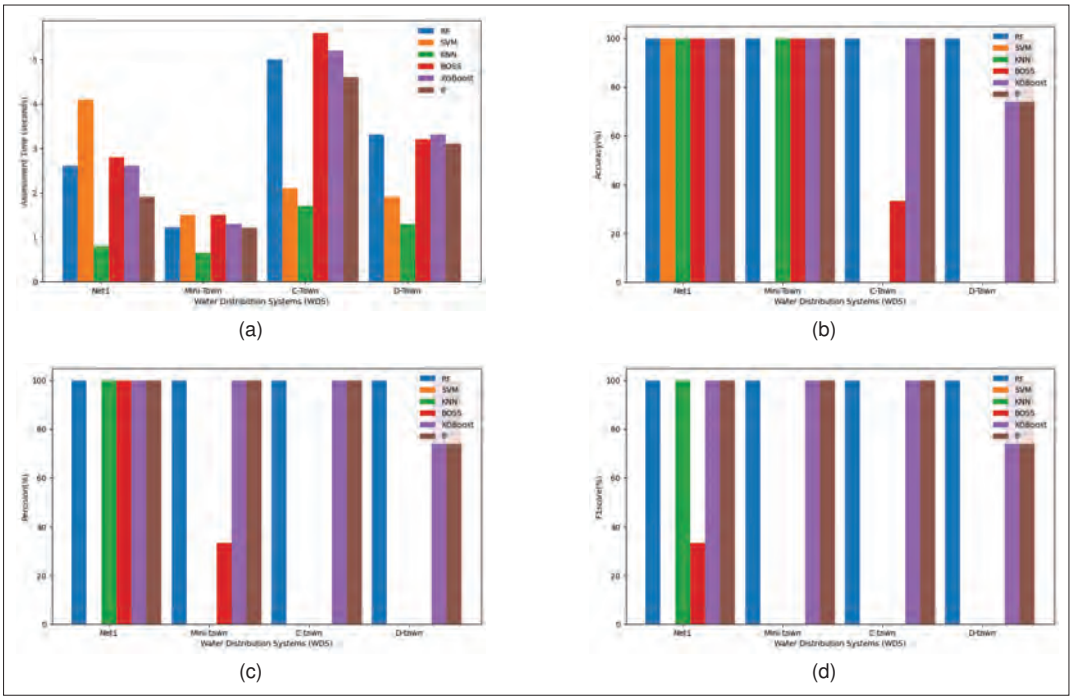


FIGURE 2. Performance evaluation of attack detection system using six classifiers (RF, SVM, KNN, Boss, XGboost and IF) on four datasets (NEt1, Mini-town, C-town and D-Town) in terms of: a) assessment time; b) accuracy; c) precision; and d) F1score.

sion attributed to its inherent limitations in handling high-dimensional data or datasets with significant overlap between classes, resulting in more false positives. KNN achieves 100% precision for Net1, demonstrating effectiveness true positives with zero false positives. RF, XGBoost, and IF achieving 100% precision for C-town and D-town, highlighting the effectiveness of the ensemble methods for detecting attacks. 100% precision indicates all detected attacks are accurate, while 0% precision means none of the detected attacks are correct. RF, XGBoost, and IF exhibit the highest F1 scores across datasets, showcasing a strong balance between precision and recall, effectively measuring a model's attack detection accuracy while considering false positives and negatives. To enhance the system's accuracy and robustness, a combination of diverse learning techniques — namely Deep Neural Networks (DNN), RF, and IF is employed in a voting mechanism. The accuracy rates for Net1, Minitown, C-town, and D-town are 100%, 100%, 100%, and 80% respectively. This approach ensures that the detection process is dependable, as it is validated by multiple learning methods.

Data Validation

This system assesses system complexity using latency and throughput. Latency is the time it takes for a transaction to be verified and irreversible, determined by block generation time (t_G) and mining time (t_m). Throughput, measured in transactions per second ($TPS = \frac{N_T}{t_L}$)., depends on the number of transactions (N_T) and latency ($t_L = t_G + t_{m,\delta}$) in which δ is the consensus mechanism. Security is paramount, and consensus mechanisms like PoW and PoT aim to ensure security by preventing majority power control (over 51% and 33.3% mining power, respectively). The formula $\nu_M \leq \nu_{T,\delta}$ helps guarantee security against malicious verifiers (ν_m), considering the number of true validators (ν_T) and total nodes (N). While PoA has the least behavior, PoW allocates all peers in parallel, resulting in longer mining times. Other water distribution systems exhibit similar behavior. The analysis presented in Table 3 investigates the Net1, Minitown, C-town, and D-town datasets, focusing on consensus mechanisms including PoW, PoT, PoA, and PoV. Each block is constructed with one measurement from each sensor, and the number of transactions varies from 336 to 953. The study calculates these scenarios' generation, mining, total time, and throughput. Notably, the results reveal that PoT and PoA achieve the highest throughput, followed by PoV and PoW, as they require only one selected validator compared to PoV and PoW, which mandate the participation of all nodes in the validation process. The throughput spans 1.35 to 8.07, depending on the dataset and the consensus mechanism employed.

Implementing blockchain at the water tank level safeguards sensing and control signals, preventing unauthorized alterations that could lead to potential flooding or drought. Some implementations in the network have failed to incorporate consensus mechanisms with an authorization point for data approval. Regarding data validation, water systems require strong security and rapid verification. To meet these criteria, PoT and PoV are proposed consensus mechanisms. PoV offers stronger security but consumes more energy and experiences delays due to data processing by all chosen blockchain peers. Given the limited number of blockchain nodes in water systems, real-time processing remains feasible. If stringent security measures are not necessary, PoT can be used to reduce resource consumption.

Using water tanks as authorized blockchain nodes addresses the significant processing require-

WDS	Cons.	Transac	Gener. time (sec)	Mining time (sec)	Total time (sec)	Throughput (TPS)	Time per Transaction (Sec/Trans.)	Trans. Efficiency (T^2/Trans.)
Net1	PoW	336	3.36	247.30	250.6643	1.34	0.746	0.00534
	PoT			48.59	51.95	6.46	0.1546	0.12435
	PoA			53.13	56.49	5.94	0.1681	0.1051
	PoV			99.50	102.86	3.26	0.30619	0.03169
Mini-Town	PoW	504	5.04	462.478	467.518	1.07	0.927615	0.002
	PoT			57.87	62.91	8.01	0.12482	0.127
	PoA			61.69	66.73	7.55	0.13240	0.1131
	PoV			260.38	265.42	1.89	0.5266	0.0071
C-Town	PoW	953	9.53	872.6	882.13	1.08	0.9256	0.00122
	PoT			109.2	118.73	8.07	0.12458	0.067969
	PoA			116.4	125.93	7.56	0.1321	0.0600
	PoV			491.3	500.83	1.90	0.5255	0.00379
D-Town	PoW	381	3.81	277.87	281.68	1.35	0.7393	0.004792
	PoT			54.6	58.41	6.52	0.1533	0.111624
	PoA			59.7	63.51	6.0	0.1666	0.09447
	PoV			111.8	115.61	3.29	0.303438	0.02845

TABLE 3. Performance Evaluation of Data Validation towards four datasets (NEt1, Mini-town, C-town and D-town) using four Consensus Mechanisms (PoW, PoT, PoA and PoV) in terms of number of transaction, generation, mining and total time, as well as throughput time per transaction and transaction efficiency.

ments for transmitting and verifying water measurement data, especially when other components of the EPANET system lack computational ability. The mining time for verifying transactions varies according to the consensus techniques used. PoA has the shortest mining time, while PoW has the longest due to its intensive parallel data verification process. The PoT and PoV consensus processes are recommended in the context of water systems due to their ability to provide comprehensive security while allowing real-time processing. These measures successfully protect the integrity and security of data. Mining, complexity, and throughput parameters were used to analyze and compare the efficiency and performance of several consensus techniques, including PoA, PoT, PoW, and PoV. These measurements provide useful information about their unique capabilities and performance characteristics.

Blockchain technology provides several security benefits in water distribution systems, such as ensuring service availability, data immutability, transparency, and preventing single points of failure. It uses mining problems to train detection models and improves data integrity by disseminating datasets and learning algorithms across many blockchain nodes. However, key differences include the lack of communication protocol simulation, the exclusion of WDS controllers (PLCs), the disregard for broadcasting delays and data transfer processing, the requirement for direct communication among selected blockchain nodes, and vulnerabilities to Man-in-the-Middle (MiTM) attacks in certain consensus methods such as PoA. These distinctions emphasize the system's unique features and limits compared to real-world blockchain implementations in water distribution networks.

Future Research Directions

The extensive research on this topic offers numerous opportunities for future work within water applications and related IoT domains. This research includes:

- Integrating communication protocol simulations with WDS controllers is crucial. Simulating communication protocols allows for the testing and optimization of consensus mechanisms under various conditions, identifying potential issues like bottlenecks, latency, and security vulnerabilities. Including WDS controllers in these simulations adds a layer of realism, as these controllers manage physical components like pumps and valves, which significantly affect system performance and reliability.
- Extending the model to include hardware implementation is another point to explore. Implementing the best-suited model specifications, selected based on this model's outcomes, can be applied to real water distribution systems. This implementation can be investigated on various platforms, including Amazon Web Services, Google Cloud, IBM Cloud, Ethereum, Hyperledger Fabric, and IBM Watson.
- Investigating methods to optimize assessment times could significantly enhance the system's overall efficiency and responsiveness. This can enhance the scalability of this system to consider multiple subsystems in the water domain. For instance, incorporating adaptive algorithms that dynamically adjust based on network load and transaction volume could mitigate delays. Moreover, exploring parallel processing techniques or decentralized validation approaches may help distribute the computational load more evenly, thereby reducing bottlenecks and improving assessment times.
- The architecture's capabilities can be broadened to advance decision-making processes through optimization. Decision-making actions range from isolating infected components to network shutdowns for recovery. Integrating other blockchain use-cases, such as peer-to-peer trading models, can be explored to promote robust and precise decision-making, contributing to the shift towards autonomous decision-making. This system can dynamically adapt to changing conditions in the network. Optimizing node selection for blockchain validators is a critical area to explore, as it can significantly impact the efficiency and security of the system. Imple-

Communication interception can allow attackers to hijack control signals, causing operational disruptions and potentially endangering public health.

menting adaptive system management strategies will ensure that the network can respond to various challenges in real-time, enhancing overall resilience and performance.

- Extending the data validation model involves constructing a system that can select the most suitable nodes to serve as blockchain nodes. This selection process should be based on specific criteria like distance and computational capabilities. Since certain sensors and IoT devices may not support validation computations, introducing a novel approach for node selection becomes crucial to keep networks of varying sizes.
- Expanding the range of attack scenarios within the EPANETCPA to significantly mitigate complex advanced attacks. For instance, pump manipulation attacks can cause incorrect water pressure levels, leading to either system damage or inadequate water supply. Sensor spoofing can introduce false data into the network, resulting in incorrect system responses such as unnecessary water treatment processes or false contamination alerts. Communication interception can allow attackers to hijack control signals, causing operational disruptions and potentially endangering public health. Moreover, attacks targeting the SCADA system can disrupt the entire water distribution process, leading to severe service outages. Malware attacks on system controllers can corrupt essential software, compromising the integrity and functionality of the water distribution system. Unauthorized access to data logs can lead to the exposure of sensitive information, which can be exploited for further attacks or to undermine public trust.
- Implementing advanced detection techniques (i.e., federated learning) can significantly enhance the ability to identify and respond to threats. These methods analyze network patterns and behaviors to spot anomalies that might indicate an attack. Federated learning uses decentralized data from various sources to train models while maintaining data privacy.
- Extending on the consensus mechanism in data validation can involve ML. ML adds to alternate mining algorithms in PoW blockchains, improving security and efficiency. GENECI uses evolutionary ML to help in consensus inference by acting as an organizer for creating ensembles [14]. This includes refining data validation processes to ensure data integrity and security. Implementing multi-layered validation mechanisms can add extra layers of protection. Moreover, ML plays an essential role in safeguarding federated learning with blockchain, acting as the foundation for authenticating transactions and guaranteeing a secure and transparent federated learning environment [15].
- Enhancing the GUI is another essential point for user-friendly and adaptive system management. A well-designed GUI can simplify complex processes, making the system more accessible to users and allowing for easier monitoring and control. Integrating these elements can lead to a more robust and intuitive system, capable of adapting to various scenarios and maintaining optimal performance.

Conclusion

In this article, a proposed WDS architecture is a comprehensive framework comprising four basic components: an EPANET-based WDS model, a data validation system based on blockchain technology, cloud storage, and system services. This architecture integrates sophisticated technology to satisfy the requirement for secure and robust water distribution systems. The EPANET paradigm simplifies the construction of water distribution networks, while the blockchain-based data validation mechanism protects the confidentiality and integrity of data transmission. Cloud storage is used for data reference and system services such as SCADA, ICS, and threat detection. The design exhibits its ability to improve the security of water distribution systems by effectively identifying attacks and safeguarding data integrity through a thorough review of ML methods. Furthermore, the study provides various consensus processes, providing insights into their effectiveness in data security and timely verification.

The proposed WDS design uses advanced technology to improve the security and resilience of water distribution systems. It combines an EPANET-based architecture, blockchain-based data validation, cloud storage, and system services to form a comprehensive framework for water infrastructure management. Examining ML algorithms and consensus methods provides vital insights into enhancing attack detection and data validation. This research helps the development of reliable and secure water distribution systems by tackling critical infrastructure concerns in the context of constantly changing technologies and threats.

Acknowledgment

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Training Networks (ITN)-IoT4Win grant agreement No. [765921].

References

[1] F. Bignell, *Riskiq Find Cybercrime Cost Organizations Just Under $1.8 Million Per Minute*, July 2021.
[2] E. Goncharov, K. Kruglov, and Y. Dashchenko, "Five ICS Cybersecurity Myths Based on Kaspersky Lab ICS Cert Experience," *Automatisierungstechnik*, vol. 67, no. 5, 2019, pp. 372–82.
[3] Q. Peng *et al.*, "Automatic Monitoring System for seed Germination Test Based on Deep Learning," *J. Electrical and Computer Engineering*, 2022.
[4] A. Bayoumi and R. McCaslin, "Internet of Things — A Predictive Maintenance Tool for General Machinery, Petrochemicals and Water Treatment," *Advanced Technologies for Sustainable Systems: Selected Contributions from the Int'l. Conf. Sustainable Vital Technologies in Engineering and Informatics*, BUE ACE1 2016, 7–9 Nov. 2016, Cairo, Egypt, Springer, 2017, pp. 137–46.
[5] C. Fathy and H. M Ali, "A Secure IoT-Based Irrigation System for Precision Agriculture Using the Expeditious Cipher," *Sensors*, vol. 23, no. 4, 2023, p. 2091.
[6] H. H. M. Mahmoud, W. Wu, and Y. Wang, "Secure Data Aggregation Mechanism for Water Distribution System Using Blockchain," *2019 25th Int'l. Conf. Automation and Computing (ICAC)*, 2019, pp. 1–6.
[7] H. Zeng *et al.*, "An IoT and Blockchain-Based Approach for the Smart Water Management System in Agriculture," *Expert Systems*, vol. 40, no. 4, 2023, p. e12892.
[8] B. K. Mohanta, S. Chedup, and M. K. Dehury, "Secure Trust Model Based on Blockchain for Internet of Things Enable Smart Agriculture," *2021 19th OITS Int'l. Conf. Information Technology (OCIT)*, 2021, pp. 410–15.
[9] H. Mahmoud, W. Wu, and M. M. Gaber, "A Timeseries Self-Supervised Learning Approach to Detection of Cyber-Physical Attacks in Water Distribution Systems," *Ener-*

gies, vol. 15, no. 3, 2022, p. 914.
[10] R. Taormina *et al.*, "Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks. *J. Water Resources Planning and Management*, vol. 144, no. 8, 2018, p. 04018048.
[11] H. H. Mahmoud, W. Wu, and Y. Wang, "Wdschain: A Toolbox for Enhancing the Security Using Blockchain Technology in Water Distribution System," *Water*, vol. 13, no. 14, 2021, p. 1944.
[12] T. Nododile and C. Nyirenda, "A Blockchain-Based Secure Data Collection Mechanism for Smart Water Meters," *2023 IST-Africa Conf. (IST-Africa)*, 2023, pp. 1–8.
[13] R. Taormina *et al.*, "A Toolbox for Assessing the Impacts of Cyber-Physical Attacks on Water Distribution Systems," *Environmental Modeling & Software*, vol. 112, 2019, pp. 46–51.
[14] Adrián Segura-Ortiz *et al.*, "Geneci: A Novel Evolutionary Machine Learning Consensus-Based Approach for the Inference of Gene Regulatory Networks," *Computers in Biology and Medicine*, vol. 155, 2023, p. 106653.
[15] H. H. M. Mahmoud, W. Wu, and Y. Wang, "Proof of Learning: Two Novel Consensus Mechanisms for Data Validation Using Blockchain Technology in Water Distribution System," *2022 27th Int'l. Conf. Automation and Computing (ICAC)*, 2022, pp. 1–5.

Biographies

Haitham Mahmoud (haitham.mahmoud@bcu.ac.uk) Haitham Mahmoud received his Ph.D. from Birmingham City University (BCU) in Electrical Engineering in 2022 and is a Marie-curie fellow on the IoT4win project at BCU between 2018 and 2021. His main research interests and expertise span a broad range of areas in intelligent networks and utilizing AI on digital systems. As a research fellow at the Future Communication research cluster, he is currently actively coordinating research activities and overseeing major projects in BCU, including Intelligent warehouse tracking, intelligent B5G/6G radio access networks, radio access network resource optimization, Cognitive radio networks, massive MIMO and future open networks. His research has been supported by H2020, UK DSIT, UKRI, NGI, BC and others in Egypt (STDF and NTI). He has been the Work Package Leader of H2020 IoT4win, Co-PI for BC TSNE with Egypt and academic supervisor for AKT.

Wenyan Wu is currently a Professor in Smart Sensor and Advanced System Engineering and the Chair in the research group of Sensor and Control with Birmingham City University, Birmingham, U.K. She received the B.Eng. and M.Eng. degrees in Electronic Engineering from Dalian University of Technology, China, obtained Ph.D. degree in water quality modeling and optimization in water distribution systems from the Harbin Institute of Technology, Harbin, China, and gained the second Ph.D. degree in Reconfigurable Virtual environment Design for virtual testing from the University of Derby, Derby, U.K. She was a Principal Investigator (PI) and a Project Coordinator for EU FP7-WatERP, EU FP7-SmartWater, and EU Horizon 2020-IoT4Win. She has extensive research experiences in smart sensors and sensor networks, the IoT, intelligent monitoring and Artificial intelligence, digital design and processing, data visualization, modeling and optimization in water distribution system, and water resource management. She has authored or coauthored more than 150 peered reviewed journal articles and conference papers. She was a Professor in Digital Design and Technologies with Staffordshire University, U.K. a Senior Lecturer in Computing, Harbin Institute of Technology, China and a Research Fellow in Water Software Systems, De Montfort University, U.K., and the National Key CAD Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. She organized special issue in IEEE journals and regular IEEE journal reviewer and was member of IEEE CIS and IEEE sensor council.

Mohamed Medhat Gaber received the Ph.D. degree in artificial intelligence from Monash University, Melbourne, Australia, in 2016. He is currently a Professor in data analytics with the School of Computing and Digital Technology, Birmingham City University, Birmingham, U.K. He then held appointments with the University of Sydney, Camperdown, CSIRO, Canberra, and Monash University, all in Australia. Prior to joining Birmingham City University, he worked for the Robert Gordon University as a Reader in computer science and with the University of Portsmouth as a Senior Lecturer in Computer Science, both in the U.K. He has authored or coauthored more than 200 papers, coauthored three monograph-style books, and edited/coedited six books on data mining and knowledge discovery. His work has attracted well more than 11000 citations, with an H-index of 45. He was in the program committees of major conferences related to data mining, including ICDM, PAKDD, ECML/PKDD and ICML. He has also co-chaired numerous scientific events on various data mining topics. He is recognized as a Fellow of the British Higher Education Academy (HEA). He is also a Member of the International Panel of Expert Advisers for the Australasian Data Mining Conferences. In 2007, he was the recipient of the CSIRO teamwork award.

Yonghao Wang is an Associate Professor with the Cyber Physical Systems research group at Birmingham City University, UK. With over two decades of experience in the fields of Networking and Cybersecurity, he specializes in blockchain technology and decentralised communication systems. His expertise extends to Digital Signal Processing (DSP) for multimedia, as well as Cloud and Edge computing. His research interests include the development of ethical and accessible blockchain technologies for inclusive finance, time-deterministic systems for beyond 5G communications, and the integration of AI/ML in network performance optimization. He is also an active contributor to international standardisation efforts, particularly through his work with ETSI and AES standards.

Blockchain-as-a-Service: Architecture, Opportunities and Challenges

Syed Muhammad Danish, Gautam Srivastava, Reza Nourmohammadi, Nouman Ashraf, Ali Ranjha, and Aroosa Hameed

Abstract

Blockchains are usually managed by blockchain nodes, which maintain a copy of all the blockchain's data and participate in validating transactions and reaching consensus with other blockchain nodes. However, running a blockchain node on your own is not easy due to the high maintenance costs and specialized hardware needed. Blockchain-as-a-service has been introduced recently by cloud giants to enable enterprises to manage blockchain nodes and networks by abstracting infrastructure setup complexities. While current BaaS solutions simplify integration and development, they suffer from inefficiencies due to fixed resources, scalability challenges, and cost inefficiencies. The purpose of this article is to analyze the integration of blockchain technology with cloud computing. In particular, we identify the costs, performance, scalability, and other challenges relating to blockchain-as-a-service. As part of our proposal, we suggest dynamic resource allocation, optimizing node computation to match web3 application requirements, and improving blockchain node scalability. The real-time adaptability of this approach ensures cost efficiency and performance improvements as workload changes. Finally, we provide research directions relevant to future research that will be required to fully utilize blockchain and cloud technology.

Introduction

Over the past decade, the Internet of Things (IoT) [1] paradigm has evolved rapidly, leading to its adoption in critical infrastructure. Although critical infrastructure systems [2] are crucial to society and the economy, the IoT paradigm benefits are short-lived due to the exponential rise in security and privacy threats associated with it. Attackers use privacy-targeted attacks to access sensitive and confidential information about critical infrastructure in pursuit of their self-interest, political ends, and commercial interests. Emerging blockchain technology [3, 4] has exhibited excellent features that can cope with the existing security and privacy issues. Blockchain, or distributed ledger, is a series of immutable transaction records, so if one of the records is modified, the rest of the peers will invalidate the transaction. Blockchain technology is trustless, meaning that it does not require third-party verification (i.e. trust), but instead uses a powerful consensus mechanism with crypto-economic incentives to verify the authenticity of a transaction in the database, which also makes it safe, even in the presence of powerful or hostile third parties trying to prevent users from participating. Apart from the financial field, Blockchain Technology has been integrated into other sectors like fraud detection, the Internet of Things (IoT), smart grids, healthcare applications, etc. A basic architecture of web3 application is shown in Fig. 1.

A blockchain is usually managed by blockchain nodes, which maintain a copy of the entire blockchain's data and participate in validating transactions and reaching consensus with other nodes. Having a blockchain node running for a Web3 application has many advantages, such as ensuring complete isolation and autonomy to protect privacy, ensuring transactions are broadcast at any time to prevent censorship, and ensuring full control over the node's software and configurations. However, running a blockchain node on your own is not easy since it requires dedicated hardware (e.g. RAM, storage, etc.) with high maintenance costs, and can have many technical and reliability issues, such as bugs in software updates, CPU spikes, and memory leaks. Web3 application developers usually use third-party RPC node providers, like Infura, Alchemy, QuickNode, etc., to avoid these issues.

In recent years, big cloud companies such as Amazon Web Services (AWS) [5] and Google Cloud Platform (GCP) [6] have introduced blockchain-as-a-service [7], which allows enterprises to deploy their nodes to deploy blockchain networks with just a few clicks, eliminating the complexity of infrastructure setup and configuration. The BaaS model allows blockchain networks to be deployed, managed, and maintained via the cloud. By using this technology, businesses are able to leverage blockchain technology without having to manage the complex infrastructure themselves. Many cloud providers offer BaaS solutions, but traditional methods often require static resource allocations, which results in inefficiencies, especially in web3 environments with dynamic workloads, thereby ensuring security, privacy, decentralization, censorship resistance and sovereignty. The price for these blockchain services is determined by how many peer nodes are used per hour, their storage per GB per month, and how many requests are made through the API.

Syed Muhammad Danish is with the Algoma University, USA; Gautam Srivastava is with Brandon University, Canada; Reza Nourmohammadi is with the University of British Columbia, Canada; Aroosa Hameed and Ali Ranjha are with the École de Technologie Supérieure, Canada; Nouman Ashraf is with Technological University Dublin, Ireland.

Digital Object Identifier: 10.1109/IOTM.001.2300199

2576-3180/24/$25.00 © 2024 IEEE

The Scope of this Article: The current blockchain-cloud infrastructure offers advantages in ease of use and integration, as existing BaaS solutions simplify blockchain development and allow seamless integration with cloud environments. Despite its advantages, it also has several disadvantages. The fixed resource model of many BaaS solutions leads to inefficiencies. Scalability challenges and cost inefficiencies often result in under or overutilization of resources. In this article, we analyze the blockchain-as-a-service architecture for private and public blockchains for possible improvements by considering the following key questions:

- How to optimize the monetary cost of hosting blockchain nodes running in a cloud virtual machines/container environment.
- How to optimize the computation power of the hosting blockchain nodes.
- How to optimize the scalability of a blockchain network considering the dynamic workload.
- What can the cloud platform provider do to improve blockchain-as-a-service for the Web3 community?

The remainder of this article goes as follows. We present the blockchain-as-a-service design, while we present the proposed improvements in the current blockchain-as-a-service design. We provide the future research directions. Finally, we conclude this article in the last section.

Blockchain-as-a-Service

Blockchain-as-a-service, the combination of cloud computing and blockchain, is an offering that allows users to leverage cloud-based solutions to build, host and manage their blockchain applications, smart contracts and functions on the blockchain network. The blockchain-as-a-service providers manage all the necessary tasks and activities to keep the infrastructure agile, operational and easily accessible. This eliminates the need for manual hardware provisioning, software configuration, and networking and security configuration. A dedicated blockchain node offers many benefits over third-party RPC node providers, e.g.,

- The API requests go to a dedicated blockchain node unlike RPC calls go over the public internet, thus ensuring the requests' privacy.
- Certain regulated industries require organizations to operate in a specific jurisdiction and control their nodes
- There is no need to manually configure the node hardware for optimal performance.
- A decentralized application communicates with a dedicated blockchain node with low latency unlike RPC calls to third-party node providers.
- The dedicated blockchain node is not shared by other applications unlike RPC node provider solutions, which leads to predictable and consistent high performance of the application.

Businesses are currently focused on establishing a development environment that will support them in exploring the potential of blockchain technology. Additionally, it can provide a wide range of operation services based on blockchain, such as search queries, transaction submission, and data analysis. These services can help developers verify their concepts and models faster. Many cloud giants are currently offering blockchain-as-a-service on their platforms, including IBM, Amazon, Google, Oracle, Kaliedo, Alibaba, and VMware blockchain, among others. There are differences between public and private blockchain architecture in terms of architecture and applications.

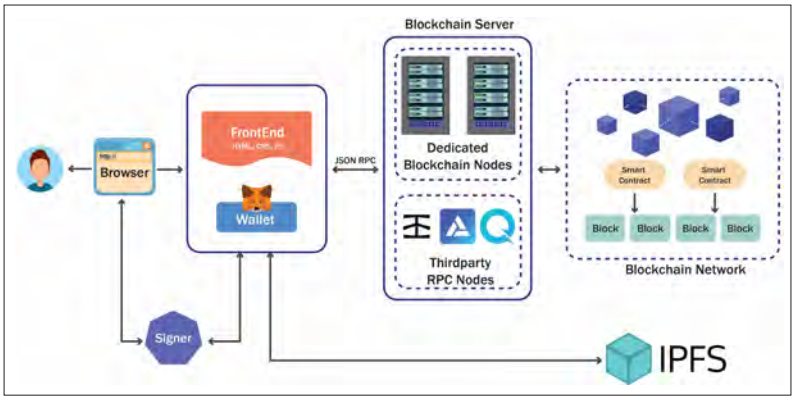


FIGURE 1. Web3 application architecture.

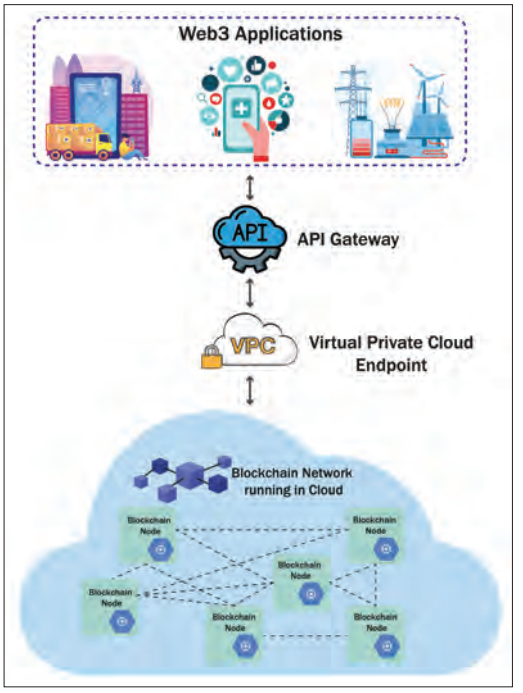


FIGURE 2. Blockchain-as-a-service private blockchain.

Using fully managed blockchain services, cloud providers can provide blockchain infrastructure for public and private blockchain [8] systems by removing the overhead associated with creating a private blockchain network or connecting to public blockchain network nodes. The enterprise can build a fully managed private blockchain in minutes with a customized consensus algorithm and blockchain parameters and then invite partner organizations to the network. Developers and organizations can use cloud services to provision a dedicated blockchain node (Ethereum, Solana) for reading data, subscribing to events, and broadcasting transactions on the blockchain. This dedicated node can then be used as the main interface to build Web3 applications on top of the public blockchain. The architecture of cloud-based private and public blockchain are shown in Fig. 2 and Fig. 3, respectively.

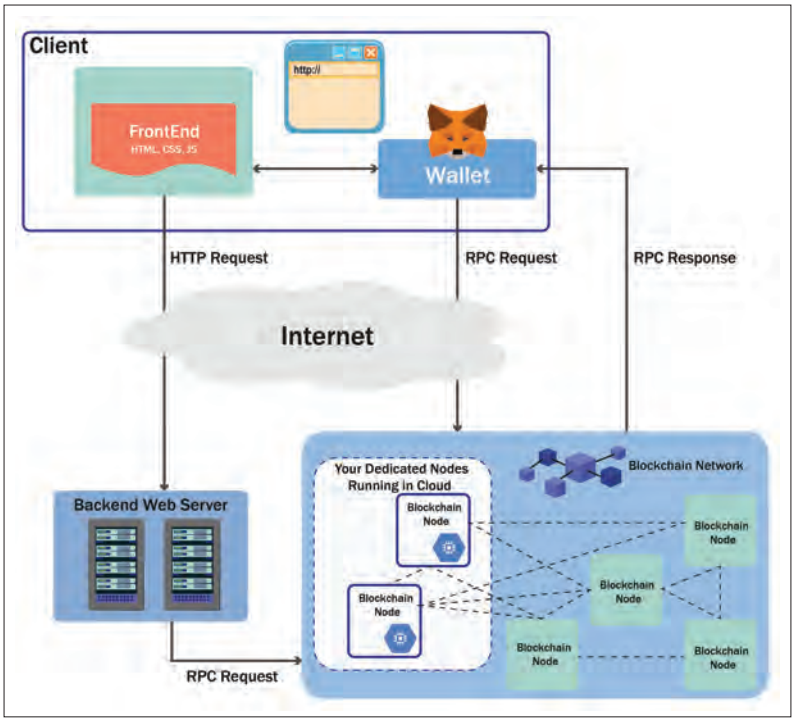


FIGURE 3. Blockchain-as-a-service public blockchain.

Current Limitations and Challenges

It is not a simple task to run blockchain nodes in a cloud premise in an optimized way. There are still some limitations related to the services provided by these platforms.

The Web3 applications [9] are designed based on their unique characteristics, which makes them different from other applications. These design patterns depend on different parameters including the application's workload. For example, a DeFi application to trade cryptocurrencies, lend or borrow assets, may experience high volumes of transactions, and is required to store a significant amount of data on-chain, thus high blockchain interaction. In contrast, the identity management web3 application might not have the same transaction volume compared to the DeFi application, and some identity proofs can be stored off-chain to lower the blockchain interaction. Considering the inherent differences and unique characteristics of Web3 applications, a proper design analysis is required to provision the cloud resources to a blockchain node. Moreover, the choice of hosting a blockchain node needs to consider the future requirements of the web3 application too, as the application might get famous and grow in the future, which could attract a large number of users. Therefore choosing the right computation power for a blockchain node is necessary, as it can lead to under-provisioning or over-provisioning of resources if not properly managed.

Cost optimization is also important, similar to computational efficiency. As described above, Web3 applications have dynamic workloads, so static resource provisioning can lead to overprovisioning when the workload is small. Due to the price of blockchain nodes based on the number of peers per hour usage, blockchain nodes running in cloud VMs may incur additional costs. Thus, monitoring the Web3 application workload over time can help optimize computation resources and associated costs by dynamically allocating resources. During periods of low workload, some idle nodes can be turned off to save money. Moreover, the price also depends on the number of requests being made to the blockchain node. Frequently sending requests to the blockchain for historical data can impact the performance and price of the blockchain node. Considering the frequency of the interaction of web3 application with the blockchain node, the cost can be optimized by carefully architecting the application design.

Moreover, the scalability of the blockchain node running in the cloud is also essential from the user's perspective, since Web3 applications become popular over time, attracting large amounts of traffic, potentially leading to high user numbers and transaction volumes. As a result of this huge workload, cloud resources associated with the blockchain node must be increased to manage the workload of the application. Dynamic resource provisioning can result in cost and computation optimization as well as scalability of Web3 applications. Furthermore, since the price of cloud-hosted nodes depends on the number of nodes, the usage of nodes, and the amount of storage, the Web3 application developer should be able to assess performance, availability, and compliance. It is therefore important to periodically verify the service level agreement (SLA) terms in a transparent and trustworthy manner.

Proposed Improvements

In this section, we propose different techniques to improve the blockchain-as-a-service running on clouds in terms of performance, cost and scalability.

Workload Monitoring of web3 Application

In the basic architecture of a web3 application, we have a front-end, where the end-user interacts with the application and a blockchain network that manages user requests, as shown in Fig. 1. Some examples of these Web3 applications are decentralized/peer-to-peer trading exchanges, identity and access management applications, decentralized social media applications, etc. These different web3 applications vary significantly in terms of design and architecture and are different in terms of their workload depending on factors such as the type of application, the size of the user base, the frequency of user interactions with the blockchain network, dependability on real-time or historical data, the complexity of smart contract logic, and the volume of data stored on the blockchain, etc.

For example, in a web3 decentralized exchange [10], where the application allows users to trade cryptocurrencies, is used by a large number of clients who are actively trading cryptocurrencies. The trading of assets is being done on the smart contract, which means each transaction needs blockchain interaction. Moreover, the application requires real-time tokens and asset information, which requires a high interaction with the blockchain network. This puts a high load on the blockchain RPC node running in the cloud as new coins are introduced regularly and may require extra cloud resources in case of a hike in application users to meet the throughput and performance requirements. In contrast, identity and access management application, e.g., Energy Web Switchboard application [11]requires users to create the decentralized identity one time on the

blockchain network and this stored information on the blockchain network can be called directly without writing anything on the blockchain, thus requiring less blockchain writing interaction. %Moreover, the number and complexity of queries vary greatly depending on the application.

Considering the above scenario, some Web3 applications may require frequent real-time updates to the blockchain, while others may primarily rely on historical data and read-only calls. Therefore, the workload and nature of Web3 applications needs to be monitored continuously to make the the choice of hosting a blockchain node. By considering these parameters, we can keep track of the volume of requests to the blockchain and the size of the user base, etc., and make necessary changes to the resources associated with the blockchain node running in the cloud. This analysis will help us in provisioning cloud resources to a blockchain node and choosing the right computation power, resulting in better utilization of cloud resources to increase performance and reduce cost.

Minimize the Interaction with Blockchain Node

The pricing of hosting a blockchain node in the cloud depends on the number of requests being made to the blockchain node. Therefore, to optimize the computational efficiency of hosting a blockchain node and the overall cost associated with the Web3 application, we need to think about how we can minimize the interaction of the Web3 application interface with the blockchain node. As we see that some web3 applications rely on historical data rather than real-time data, so we need to think about how can we control the number of requests or offload them to fulfill users' requests without directly requesting resources from the blockchain node? One solution to this problem is to store the blockchain data in a high-performance database for better read performance. The historical data queries from old blocks in the blockchain are made to the database instead of the blockchain. AstraBlock is an example of such a database, which converts the blockchain data to high-performance query databases. This leads to less load on the blockchain node by reducing the number of requests, thus increasing computational efficiency and cost savings. However, it comes at the cost of managing a database. Figure 4 shows the architecture of the Web3 caching layer.

Apart from the high-performance query database, memory caching can also be used to offload blockchain queries by running a caching layer in a Docker container, such as Redis and PostgreSQL [12]. This is suitable for applications where some blockchain data does not change frequently, and applications mostly rely on historical data. One of the prime examples of these types of applications is the Energy Web Switchboard web3 application, which allows organizations to manage the decentralized identifiers and verifiable claims associated with the users of their organization. This application includes the role definition and role issuers. A certain role is issued to the user one time on the blockchain network, and the user can show proof of the claim/role every time they want to perform some action inside the organization. Moreover, they also implement a back-end server, which keeps track of these decentralized identifiers and verifiable claims in the Redis memory database. Using

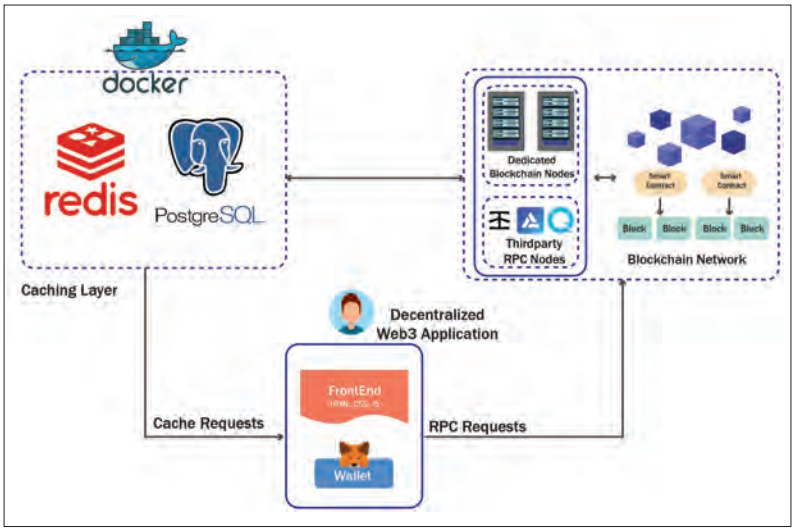


FIGURE 4. Web3 caching layer.

this design, the application does not call the smart contract functions on the blockchain node every time they need some information about the identity or verifiability of the claims, as the information can be kept in an in-memory database. The benefits of minimizing the interaction with the blockchain node are twofold, First, it helps offload the requests to the database, which increases the performance of Web3 applications and makes the blockchain node computationally efficient with less load. Second, if the blockchain node is running on a resource-based pricing model, this can reduce costs since the user's applications do not directly query or interact with the blockchain nodes; rather, they receive the same data from up-to-date databases.

Predicting the Application's Workload

As discussed in the previous subsection the web3 application's workload is dynamic and needs continuous monitoring to actively handle the provision of resources to the blockchain node. We can think of it as a dynamic workload scenario where the blockchain nodes are not always serving the users' requests. If we have a prediction of the future workload of the application along with continuous monitoring, we can intelligently allot sufficient cloud resources to the blockchain node. The basic architecture of application load monitoring and prediction is shown in Fig. 5.

To realize this scenario, Neural networks (NN) [13] can be effectively employed to predict application parameters and facilitate cost and performance optimization by accurately forecasting application workload. By analyzing historical data and training the neural network model, it becomes possible to make predictions about future workload patterns. The neural network model learns the underlying relationships between various input features such as time of day, user interactions, and historical workload data, and the corresponding application parameters like CPU usage, memory consumption, and network traffic. Once trained, the model can generalize and make accurate predictions about future workloads based on new input data. These predictions enable proactive resource allocation and optimization of cloud VMs, ensuring that the right amount of resources is provisioned at any given time.

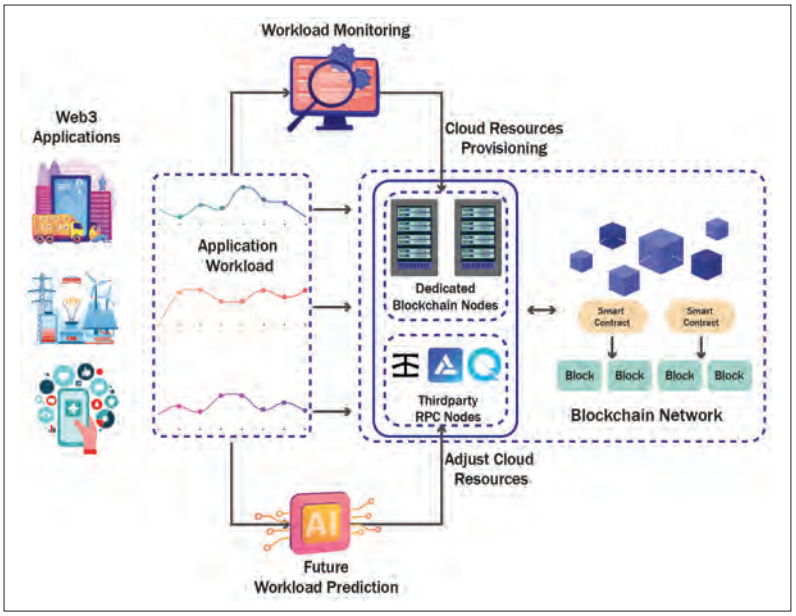


FIGURE 5. Application workload monitoring.

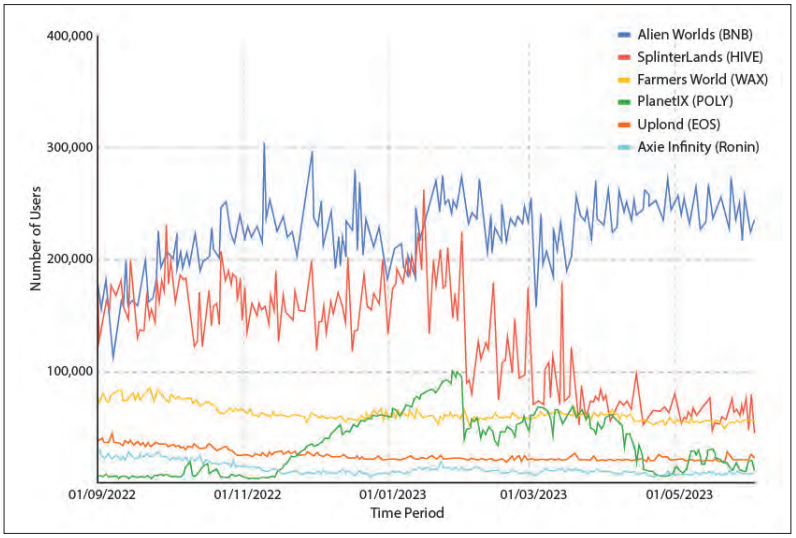


FIGURE 6. Comparison of active number of users.

With the predicted workload information, cloud VM resources can be dynamically adjusted to match the anticipated demands. For example, if the neural network predicts a surge in application workload during peak hours, additional VM instances can be provisioned in advance to handle the increased traffic. Conversely, during periods of low workload, unnecessary VM instances can be scaled down or even shut down temporarily to reduce costs. Another solution could be to allow developers to run multiple blockchain nodes rather than a single blockchain node with high computational power, and use a load balancer between them. As the workload is dynamic, and there can be times when the blockchain nodes are not receiving many requests, we can turn the idle blockchain nodes off. This brings two benefits: firstly, we can avoid overprovisioning of resources, and secondly, we can dynamically optimize the computation power of blockchain nodes. We can also use auto-scaling policies that automatically adjust the number of blockchain nodes.

Figure 6 shows the different numbers of users for different blockchain-based games over some time. It can be seen that the games implemented in different blockchain networks have different numbers of users, and the number changes continuously over some time, depending on the popularity of games and other factors including performance, transaction fee, etc. Different number of active users results in different application workload on different blockchain networks, and the workload changes with the number of blockchain users. This analysis emphasizes the importance of workload monitoring as this information can optimize computation power as well as the scalability of the blockchain nodes running in the cloud.

Scalability

Blockchain nodes running in a cloud environment have different scalability requirements for private and public blockchains. As an example, Ethereum's block time is approximately 12 seconds, which means a block will be mined every 12 seconds, in a public blockchain. Despite introducing many blockchain nodes to the Ethereum network, we cannot increase the transaction rate or improve block time. Furthermore, we cannot change the inherent consensus algorithm to make it more scalable. In the case of a public blockchain node deployed in the cloud, we cannot control the write requests since they are processed according to the rules of the public blockchain. Furthermore, if we run validator nodes of the public blockchain, these nodes must run 24/7 to validate blocks and transactions because the blockchain network produces blocks constantly. Unlike public blockchains, private blockchains allow us to control the associated parameters, as well as the consensus mechanism to control scalability. We can develop our consensus algorithm or create multiple blockchain nodes based on the needs of Web3 applications if we wish to process transactions more quickly.

To increase the scalability of private blockchain inside a cloud environment, we have different techniques that can be used, such as sharding [14]. Sharding is a technique that allows a blockchain network to scale by partitioning the network into smaller pieces, or shards, and processing transactions in parallel on each shard. This can greatly increase the capacity of the network and reduce transaction times. Considering the case of a private blockchain, where the number of nodes and participants is limited and known, sharding can be a viable way to improve performance and scalability. We can define multiple blockchain nodes in the cloud VMs and by dividing these nodes into smaller networks, we can create more manageable shards. Each shard can handle a subset of the network's transactions, reducing the load on any individual node or shard. This can help the private blockchain to scale more effectively and handle more transactions per second. We can also introduce the concept of dynamic sharding [15[, where we can monitor the transactions being generated by the web3 application and change the number of blockchain nodes and shards to ensure computational optimization. In addition, consensus algorithm optimization is another way to increase the transaction throughput, leading to improvements in the speed and scalability of the network. In private blockchains, we can tune the consensus algorithms to achieve this. For example,

by reducing the block time or block size, we can process transactions faster in the blockchain network. Finally, cost-effective scalability is possible by carefully considering the application dynamics and workload and scaling the resources up and down to fulfill the application requirements.

Future Research Directions

Blockchain-as-a-service is still in its infancy, and several critical challenges need to be addressed before the potential of using blockchain in the cloud can be fully realized. Some of the leading research challenges are discussed as follows.

Some blockchain applications require real-time communication and are time-sensitive, so the blockchain nodes running in the cloud must comply with the service level agreement (SLA) and quality of service (QoS) requirements. However, we cannot rely on third-party monitoring tools to monitor cloud resources running blockchain nodes. Additionally, as the price of these nodes depends on the number of requests made via API calls and the number of peer nodes per hour, a blockchain-based traceable and auditable QoS monitoring system could be a worthwhile research effort. Moreover, deep learning and artificial intelligence techniques can be applied to analyze the workload of web3 applications to optimize the performance and cost of blockchain nodes, for example, time series-based recurrent neural networks (RNNs) can be used to predict application workload in the future to intelligently select the number of blockchain nodes in the cloud. One possible direction of research could be the allocation of cloud resources dynamically to optimize the performance and cost of Web3 applications with dynamic workloads.

Furthermore, to secure important business information, the security and privacy aspects of running a blockchain node in a cloud premise should also be considered. Data integrity, confidentiality, and availability are crucial to Blockchain-as-a-Service (BaaS). Among the key areas of concern is IoT integration, in which devices collect and transmit large amounts of data to blockchain networks. The limited computational power and security features of IoT devices make them potential targets for attacks. For example. blockchain transactions could be tampered with or unauthorized access could be gained through insecure IoT devices connected to a BaaS platform. In order to ensure the reliability and safety of BaaS environments, robust security measures need to be implemented that address the unique vulnerabilities of IoT devices. Furthermore, cloud-based blockchain nodes must be scalable given the high demand for web3 applications. Finally, layer2 solutions, side-chains, optimistic and zk-rollups could be research directions for improving the scalability of a public blockchain network.

Conclusion

The purpose of this article is to examine the integration of blockchain technology with cloud computing. Technology is still in its infancy, and several critical issues must be addressed. The first step was to identify the costs, performance, scalability, and other challenges associated with integrating blockchain technology with cloud computing. After that, we proposed several enhancement techniques to improve the performance, scalability, and cost of blockchain nodes. In Web3 applications with varying computation and cost requirements, using different monitoring tools and databases can significantly improve the performance and scalability of blockchain applications. Finally, we discuss future research directions aimed at addressing the critical challenges associated with integrating blockchain and cloud technologies to maximize cloud's potential.

References

[1] S. M. Danish, K. Zhang, and H.-A. Jacobsen, "Blockaim: A Neural Network-Based Intelligent Middleware for Large-Scale IoT Data Placement Decisions," *IEEE Trans. Mobile Computing*, vol. 22, no. 1, 2021, pp. 84–99.

[2] Y. Jiang *et al.*, "Lblockchaine: A Lightweight Blockchain for Edge IoT-Enabled Maritime Transportation Systems," *IEEE Trans. Intelligent Transportation Systems*, vol. 24, no. 2, 2022, pp. 2307–21.

[3] B. Li *et al.*, "Scenarios Analysis and Performance Assessment of Blockchain Integrated in 6G Scenarios," *Science China Information Sciences*, vol. 67, no. 7, 2024, p. 170301.

[4] Y. Li, X. Luo, W. Zhao, and H. Gao, "Reputation-Based Stable Blockchain Sharding Scheme for Smart Cities with IoT Consumer Electronics: A Deep Reinforcement Learning Approach," *IEEE Trans. Consumer Electronics*, 2024.

[5] A. E. C. Cloud, "Amazon Web Services," Retrieved November, vol. 9, no. 2011, 2011, p. 2011.

[6] J. J. J. Geewax, *Google Cloud Platform in Action*, Simon and Schuster, 2018.

[7] J. Singh and J. D. Michels, "Blockchain as A Service (baas): Providers and Trust," *2018 IEEE European Symp. Security and Privacy Workshops (EuroS&PW)*, 2018, pp. 67–74.

[8] D. Guegan, "Public Blockchain Versus Private Blockhain," 2017.

[9] D. Sheridan *et al.*, "Web3 Challenges and Opportunities for the Market," arXiv preprint arXiv:2209.02446, 2022.

[10] S. Malamud and M. Rostek, "Decentralized Exchange," *American Economic Review*, vol. 107, no. 11, 2017, pp. 3320–62.

[11] S. Hartnett *et al.*, "The Energy Web Chain-Accelerating the Energy Transition with an Open-Source, Decentralized Blockchain Platform," *Energy Web Foundation*, 2018.

[12] A. I. Sanka, M. H. Chowdhury, and R. C. Cheung, "Efficient High-Performance FPGA-Redis Hybrid Nosql Caching System for Blockchain Scalability," *Computer Communications*, vol. 169, 2021, pp. 81–91.

[13] B. Müller, J. Reinhardt, and M. T. Strickland, *Neural Networks: an Introduction*, Springer Science & Business Media, 1995.

[14] P. Zheng *et al.*, "Aeolus: Distributed Execution of Permissioned Blockchain Transactions via State Sharding," *IEEE Trans. Industrial Informatics*, vol. 18, no. 12, 2022, pp. 9227–38.

[15] Z. Yang *et al.*, "Sharded Blockchain for Collaborative Computing in the Internet of Things: Combined of Dynamic Clustering and Deep Reinforcement Learning Approach," *IEEE Internet of Things J.*, vol. 9, no. 17, 2022, pp. 16,494–16509.

Biographies

Syed Muhammad Danish (syed.danish@algomau.ca) is currently employed as an assistant professor at Algoma University, Brampton, Canada. He received his Ph.D. from ETS Montreal, Quebec.

Gautam Srivastava (srivastavag@brandonu.ca) is currently a Full Professor in the Department of Math and Computer Science at Brandon University, Canada.

Reza Nourmohammadi (reza.nourmohammadi@ubc.ca) was a Post-Doctoral Fellow at Blockchain@UBC. He earned his Ph.D. in Artificial Intelligence at the École de technologie supérieure Montreal.

Nouman Ashraf (nouman.ashraf@tudublin.ie) received the Ph.D. degree in electrical engineering from Frederick University, Cyprus, under the Erasmus Mundus Scholarship Program. He was with the Turku University of Applied Sciences, Finland, TSSG, Waterford Institute of Technology, Ireland, and the University of Cyprus. He is currently with Technological University Dublin, Ireland.

Ali Ranjha (ali-nawaz.ranjha.1@ens.etsmtl.ca) received the Ph.D. degree in engineering from École de Technologie Supérieure (ÉTS), Université du Québec, Montréal, Canada. He is currently a Postdoctoral Researcher with ÉTS, Université du Québec.

Aroosa Hameed (aroosa.hameed.1@ens.etsmtl.ca) is currently an Ericsson Postdoctoral Fellow at Carleton University, Ottawa, Canada. She received the Ph.D. degree from École de technologie Supérieure (ETS), Universite du Quebec, Montreal, Canada.

Connected Internet of Things for Monitoring and Tracking of Endangered Whales

Rodolfo W. L. Coutinho and Azzedine Boukerche

Abstract

Connected Internet of Things (CIoT) integrates Internet of Things (IoT) in different domains (e.g., spatial, aerial, terrestrial, and underwater). CIoT enables monitoring and tracking in remote and large geographic areas, such as the Earth's poles, forests, and oceans. In this article, we envision a CIoT system for the near-real-time monitoring of endangered whale species. In the envisioned system, very high-resolution images from satellites and Internet of CubeSats shall be used for the autonomous detection and location determination of endangered whales. The obtained locations shall be used to determine the trajectory of surface autonomous vessels that will temporarily deploy an Internet of aerial and underwater things for the near-real-time monitoring of detected whales. We discuss the main entities involved in the envisioned architecture, data flows, and communication paradigms needed to implement the proposed CIoT architecture and the challenges to empower the envisioned system. We also point out future research directions to be solved towards a CIoT system for whale monitoring.

Introduction

Connected Internet of Things (CIoT) has gained increased attention as the interconnection of "things" in different domains, i.e., space, air, ground, and underwater, will unlock next-generation smart applications. Internet of Space Things through CubeSats, combined with the Internet of Aerial, Terrestrial, and Underwater Things, will provide global connectivity at low costs to IoT aerial, ground, and underwater IoT devices [1]. CIoT will empower large-scale monitoring and reconnaissance applications, machine-to-machine (M2M) applications in remote areas, surveillance, ocean of things, and deep-space applications. One key characteristic of connected IoT is continuous connectivity for IoT devices located in remote and large geographical areas, such as deep space, deep ocean, and polar regions.

Connected IoT emerges as a key enabler technology for large-scale monitoring and tracking of Earth's remote areas. CIoT shall have a groundbreaking role in supporting the monitoring of endangered whales, such as the North Atlantic Right Whales. The real-time detection, monitoring, and tracking of whales is essential for preserving endangered whales. North Atlantic Right Whales (NARWs) are at risk of extinction. Its current population is less than 400 individuals. While whaling was the principal source of risk for NARWs in the past, ship collisions and entanglement in fishing nets are the main threats endangering NARWs today. The NARWs seasonally migrate through routes that overlap with busier shipping lanes and are close to major ports. This makes them more susceptible to vessel strikes and entanglements in fishing nets. Canada implements temporary ship speed restrictions from April to November, in areas where NARWs are commonly present. Moreover, time-space restrictions of fixed gears fishing are employed to reduce the chances of whale entanglement on fishing gears. The effectiveness of such approaches depends on the success of detecting NARWs.

However, the detection and monitoring of NARWs are challenging. NARWs spend over 80% of their time underwater and can be highly mobile. Traditional approaches to detect and monitor NARWs include aerial surveys, observation missions boats and high cliffs, and passive acoustic monitoring, which are often susceptible to weather conditions and have limited coverage. Instruments, such as infrared cameras on vessels, hydrophones, gliders, and drones, are being considered for monitoring NARWs [2]. However, they have limited sensing coverage in remote and large geographic areas.

This article aims to:

1. Enhance the community's awareness of the fact that NARWs are close to extinction and the need and challenges involved in the preservation of this whale species
2. Provide a roadmap for the development of CIoT for NARW monitoring

We discuss an envisioned connected IoT architecture for the large-scale, near-real-time detection, monitoring, and tracking of NARWs. The contributions of this article are the following:

- A thorough revision of the limitations and challenges encountered in the current approaches often used for the detection and monitoring of NARWs;
- The design of an envisioned architecture for connected IoT for the detection and monitoring of NARWs;
- The discussion of main entities involved in the envisioned system, and the data flows and communication paradigms needed to

Digital Object Identifier: 10.1109/IOTM.001.2400002

Rodolfo W. L. Coutinho is with Concordia University, Canada; Azzedine Boukerche is with the University of Ottawa, Canada.

2576-3180/24/$25.00 © 2024 IEEE

implement the proposed CIoT architecture;
- The future research directions to be addressed to empower the proposed CIoT architecture for monitoring NARWs.

Current Solutions for NARW Monitoring

This section discusses the working principle, advantages, and disadvantages of current approaches used to detect NARWs.

Aerial Surveillance: Aerial surveillance is the most common approach for detecting NARWs. Observers at an aircraft search for whales at the ocean surface. The aircraft flies in a grid-like pattern over the area of interest. When a whale is spotted, observers record the position, number of detected whales, dive times, ocean conditions, and evidence of entanglements or injuries. Moreover, observers take photos to identify the NARW's individuals by their callosities and unique markings. Aerial surveillance is used during the seasonal NARW migration to understand the distribution and abundance of NARWs. Aerial surveillance is expensive and limited by daylight and weather conditions (e.g., fog, rain, snow, and high winds).

Acoustic Surveillance: Acoustic surveillance uses arrays of hydrophones moored at a location or attached to mobile platforms, such as gliders, profiling floats, and surface drifters, which will follow a pre-defined trajectory in the areas of interest. The hydrophones detect and record specific sound frequencies. The advantage of passive acoustic surveillance is that it can continuously monitor marine mammals in the surrounding area from the sound they produce. Moreover, such monitoring can last long periods and cover tens to thousands of kilometers. However, PAM systems have their spatial and temporal performance affected by human-made noise, such as shipping traffic.

At-Sea Vessels Visual Survey: At-sea vessel surveys can happen from opportunistic sightings reported by vessels and boats or through at-sea vessel survey missions. In at-sea vessel survey missions, infrared cameras at vessels identify whales at the sea surface by measuring temperature differences between the animals and the surrounding water and air. Infrared cameras deployed on at-sea vessels work independently of daylight and present good performance even in distinct environmental conditions. However, it might misclassify birds and boats as marine mammals.

Underwater Sensor Networks for Ocean Monitoring: Trends and Limitations

Underwater wireless sensor networks (UWSNs) have been proposed for the autonomous monitoring of aquatic environments. This approach consists of the deployment of underwater sensor nodes that gather data of interest and collaboratively report collected data to a monitoring center.

Ocean Networks Canada (ONC) has deployed underwater physical instruments across Canada's coastal areas and launched the Ocean 3.0 Data portal, where researchers can download and visualize collected data. The physical instruments (e.g., bottom pressure recorder, hydrophone array, oceanographic instrument, acoustic imaging instrument, underwater camera, and seismometer) are anchored at the ocean bottom and connected to a collector node through fiber-optic cables. In the context of endangered whale monitoring, anchored instruments connected through cables would be feasible to monitor an area where the whales are found when feeding and nursing. This approach renders it unfeasible to monitor the whales while migrating overseas.

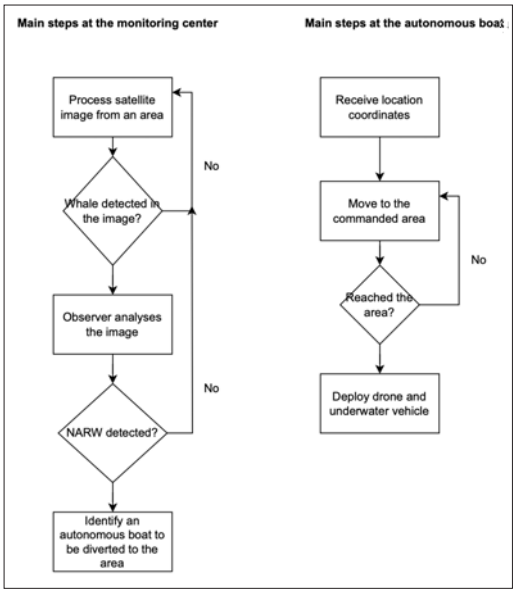


FIGURE 1. Main steps for detecting and monitoring NARWs.

Internet of Underwater Things (IoUTs) uses static and mobile underwater sensor nodes capable of communicating among themselves through wireless underwater communication. The advantages of this technology are its relatively easy deployment since there is no need for cables to connect nodes, and its self-organizing and self-configuring capabilities, where IoUT nodes work collaboratively to maintain the network and gather data from the environment. One of the challenges of this technology is mobility management since nodes might move further apart from each other due to ocean currents, resulting in network partitions and disconnections. Therefore, this technology is suitable for monitoring whales, once they have been detected, for a short period, i.e., until network connectivity lasts.

The Envisioned CIoT for Monitoring Endangered Whales

Traditional technologies for aquatic monitoring present limitations for whale monitoring scenarios. The fundamental principle of the envisioned CIoT system is that it shall provide the mechanism to monitor whales not only in feeding and nursery areas but also when migrating among areas. The envisioned system shall be able to detect NARWs, monitor NARW when migrating for as long as possible, and monitor NARW when feeding and nursing their calves. Fig. 1 shows the main requirements for the envisioned CIoT system. This section discusses our vision for a CIoT system for monitoring endangered whales. We discuss the proposed system from two main points of view:
- The different "things" used for detecting and monitoring the whales
- The communication used to integrate the different IoT systems

One of the challenges of this technology is mobility management since nodes might move further apart from each other due to ocean currents, resulting in network partitions and disconnections.

One of the challenges for whale detection from satellite imagery is the lack of datasets for training machine learning models.

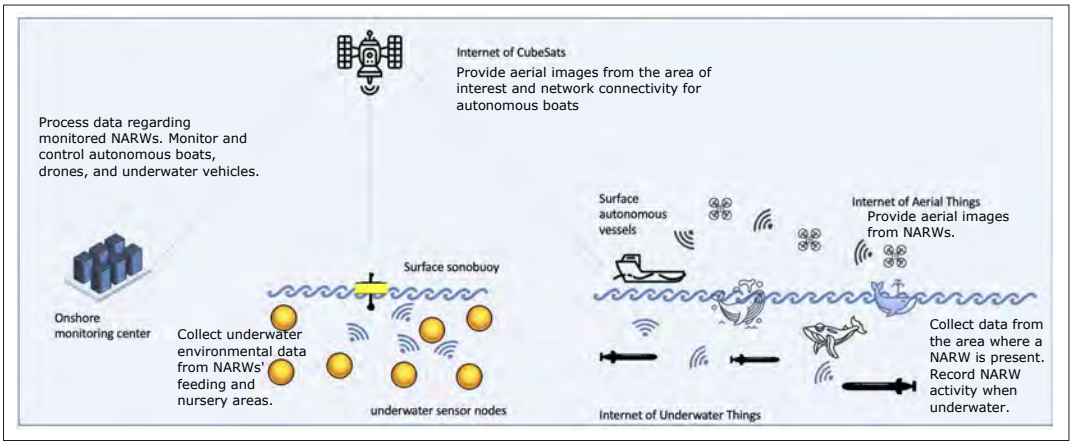


FIGURE 2. Envisioned CIoT for NARWs monitoring.

System	Devices	Functionality
Internet of Space Things	CubeSats	• Provide very high-resolution imagery of oceanic areas. • Provide communication coverage to surface autonomous vessels.
Internet of Aerial Things	Drones and unmanned aerial vehicles (UAVs)	• Used for monitoring of NARW when they surface. • Collect whale's exhaled condensed breath. • Provide aerial footage, which will be used to identify the whale from the callosities.
Internet of Underwater Things	Autonomous underwater vehicles (AUVs)	• Used for underwater data collection. • Record NARW when underwater. • Collect data regarding the area in which the NARW is present.

TABLE 1. Connected IoT for NARW monitoring.

The Envisioned Architecture: The Things

The "things" of the envisioned CIoT system are shown in Fig. 2 and summarized in Table 1. A comprehensive CIoT system for NARW detection and monitoring will comprise of Internet of Space Things, Internet of Aerial Things, Internet of Underwater Things, and autonomous vessels and underwater vehicles. The key role of each of these IoT subsystems within the CIoT system for NARW monitoring is discussed in the following sections. Figure 1 summarizes the main steps in the envisioned CIoT system for detecting and monitoring NARWs.

Internet of Space Things: Internet of CubeSats is composed of a network of small satellites that will provide wide coverage in terms of communication and imagery. CubeSats instruments will take very high-resolution images from ocean areas, which can be used to detect NARWs. For instance, DigitalGlobe (http://worldview3.digitalglobe.com) launched the WorldView-3 satellite on August 13, 2014. This satellite has a WorldView-110 camera (WV110) that provides image at 31 cm panchromatic and 1.24 m multispectral, 3.7 m in the eight short-wave infrared bands and a 30 m resolution in the clouds, aerosols, vapors, ice, and snow (CAVIS) bands (https://earth.esa.int/eogateway/missions/worldview-3).

Hodul *et al.* [3] demonstrated the feasibility of identifying NARW individuals from satellite imagery. The authors used Maxar's 15 cm resolution HD images, derived from 30 cm resolution images from the WorldView platform, taken from the area of Cape Cody Bay on April 24, 2021. A trained observer used the QGIS software and scanned the obtained image to flag potential NARWs. This resulting information was compared to information obtained from aerial surveillance of the same area. The main conclusion is that NARWs can be recognized from unique scarring and callosity patterns in satellite imagery if image resolution continues to improve.

Kapoor *et al.* [14] proposed a deep learning-based methodology for automatic whale detection from very high-resolution satellite imagery. First, areas where whales are often seen were monitored for the presence of whales. The results were used to create a dataset with manually labeled images. The images in the dataset were preprocessed to increase their quality for training the model. Hence, a modified Tiny YOLOv3 is trained and used for whale detection. The proposed model showed promising results regarding whale detection and counting.

One of the challenges for whale detection from satellite imagery is the lack of datasets for training machine learning models. Gaur *et al.* [5] developed the *SeaDroneSim2* benchmark suite to generate aerial and satellite synthetic image datasets to enhance whale detection. The SeaDroneSim2 uses Blender™ game engine to simulate the marine environment and incorporate whales. The developed tool then produces a training dataset with corresponding ground truth masks. Results obtained from the model trained with the synthetic dataset were compared with results from the model trained with a dataset of real images and demonstrated similar intersection over union performance.

Internet of CubeSats has also been designed to provide communication coverage to devices in remote regions [1]. In the envisioned CIoT sys-

tem, this technology will provide communication coverage for interconnecting autonomous boats. As will be discussed later, the envisioned CIoT shall deploy autonomous boats to carry underwater sensing vehicles. The underwater sensing vehicles will be deployed on demand, whenever a NARW is encountered in the area. They will accompany a migrating NARW for a time duration, and collect data as much as possible.

Internet of Autonomous Boats and Aerial Things: Autonomous boats and small ships are fundamental in the envisioned connected IoT system. Autonomous ships rely on machinery for decision-making and operation, i.e., port approach, departure, and ocean transit. Autonomous, zero-emission, and electric-powered ships and surface boats have been developed for military surveillance, ocean exploration, and cargo transportation (e.g., YARA Birkeland and ReVolt ships). Autonomous boats have already been designed and developed for marine monitoring missions (e.g., the F-Boat [6]).

In a CIoT, autonomous small surface boats will be used for at-sea surface mission operations for data collection of NARWs. Each autonomous surface boat will carry drones and autonomous underwater vehicles (AUVs) for aerial and underwater data collection from detected NARWs. The drones will be used for aerial footage of NARWs. The images will help identify the individual NARWs through their unique callosities, and entanglement scars, and estimate the size. Moreover, such as in the SnotBot project [7], drones can carry petri dishes to collect the whale snot, i.e., exhaled breath condensate, to determine the whale's health, diet, sex, pregnancy status, microbiome, and details about its genetics.

The AUVs launched from autonomous small boats shall self-organize themselves in a flying ad hoc network to maintain multi-hop connectivity with the surface boat. Some AUVs will stream aerial footage of the NARW to the surface boat, which will instantly transmit the received video feed to onshore monitoring centers through wireless links provided by the Internet of CubeSats. The use of underwater vehicles and drones on vessels, ships, and autonomous boats is not something new. Drones and AUVs have been proposed for autonomous inspections on confined spaces of marine vessels, for instance, due to the challenges and health risks to human surveyors [8]. Such entities already exist on vessels, ships, and boats, and they can be launched when needed for recording NARW's activities at sea.

In the mentioned CIoT system, many challenges are involved and must be tackled. First, AUVs must be capable of autonomously detecting and following NARWs. This can be done through well-trained deep learning models for NARW detection and recognition. Besides, topology control and connectivity management algorithms must be implemented to guarantee that the Internet of aerial vehicles does not lose connectivity with the surface autonomous boat.

In addition, a set of AUVs will be attached to the surface boats and deployed whenever an NARW is in the area. The AUVs will be released upon notification that a NARW is in the vicinity area. The AUVs will be in charge of acquiring data in the underwater environment near the identified NARW. The idea is to collect whale-up calls and data from chemical and physical underwater variables. They will use multi-hop underwater communication for data delivery from themselves to the surface autonomous boats. Again, the boats will use satellite communication to deliver gathered data to the onshore monitoring center.

Internet of Underwater Things: Internet of Underwater Things (IoUTs) has gained increased attention thanks to its potential to unlock smart ocean applications. IoUTs consist of heterogeneous underwater sensor nodes and autonomous vehicles (AUVs), deployed in an area of interest, which collaboratively gather data from events of interest and report them to an onshore monitoring center through satellite communication. Each underwater sensor node is equipped with an acoustic modem, which enables them to wirelessly communicate underwater.

In the envisioned architecture, underwater sensor nodes equipped with hydrophones and other sensor devices (e.g., IMSTL for measuring water temperature and depth, SBE 18 pH, HydroCAT-EP, and SeaOwl UV-A) will be used for the periodic monitoring of variables of interest in the areas of NARW are commonly seen (e.g., Shediac Valley). Those nodes will periodically collect data of interest, such as NARWs upcalls, water temperature, pH, dissolved oxygen, turbidity, chlorophyll, backscattering, and fluorescent dissolved organic matter, which will help to understand the environment where NARWs are encountered. Gathered data will be delivered, through multi-hop underwater communication, to sonobuoys or mobile autonomous boats (e.g., Teledyne Z-Boat 1800T) deployed at the sea surface. The surface nodes will offload obtained data to an onshore monitoring center through satellite communication.

Herein, the daunting challenges concern efficient and reliable data delivery, as the underwater acoustic channel is energy-hungry and presents low and variable quality. The current approach to tackle such challenges is opportunistic routing (OR) protocols. In the OR paradigm, each node with a data packet to send selects a subset of its neighboring nodes, instead of a single neighbor, as the next-hop nodes. The next-hop nodes will work in a coordinated way to forward the received packet towards the destination. Thus, a data packet is lost in a given hop only if none of the selected next-hop nodes receive it.

The Envisioned Architecture: The Data Flow

This section discusses the data flow among the devices as well as the main tasks performed by each entity. Figure 3 shows a summary of the data flow of the envisioned CIoT system for whale monitoring.

The first step is the pipelined processing of aerial images produced by the Internet of CubeSats instruments. VHR imagery from NARWs' migration routes and areas where they are commonly present will be processed by deep learning models. The model will flag possible whales in the image. When the model flags whales in an image, an experienced observer will analyze the information and identify if the whales are from the monitored species (i.e., NARW). A critical challenge in this building block consists of training a suitable ML model for detecting NARWs and confounding features, such as rocks and boats).

In a CIoT, autonomous small surface boats will be used for at-sea surface mission operations for data collection of NARWs.

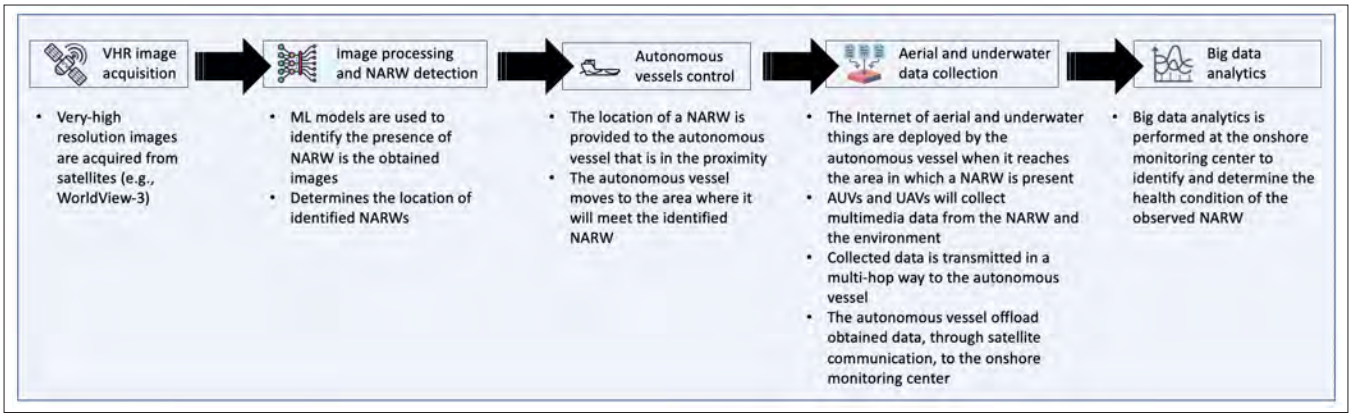


FIGURE 3. Data flow and major operations in the envisioned CIoT for NARWs monitoring.

When NARWs are detected, surface autonomous boats will be directed to the area. We envision a scenario where surface autonomous boats are strategically deployed along the NARWs' seasonal migration routes and close to ports. Each autonomous boat will be equipped with solar panels to ensure electricity. They will also be equipped with radio frequency modems for communicating with the onshore monitoring center through CubeSats. The observer will divert the autonomous boat closest to the area where a NARW has been detected, to deploy the drone and underwater vehicle for data collection. Araújo *et al.* [9] proposed the OpenBoat architecture for automating a sailboat operation for autonomous monitoring of the aquatic environment. The proximity network layer is one of the components designed in the OpenBoat architecture. Each autonomous sailboat has a 433 MHz RF antenna to enable direct communication between the sailboat and the ground station. At the monitoring center, the Mission Planner software shows the status of the sailboat sensors in a real-time manner, and it also manually or autonomously controls the boat through commands sent regarding direction, speed, and waypoints to be reached. Autonomous boats can also be connected to the onshore stations through LoRa transceivers and satellite links [10].

Each autonomous boat will also carry a drone and an underwater vehicle to be deployed near a NARW. Such devices will collect data from the NARW, such as aerial and underwater images. The underwater vehicles and drones would start collecting data about the NARW, upon they detect an individual in their coverage area. AUVs will use multi-hop communication for video streaming to the vessel. Moreover, some drones will collect exhaled breath condensate to be used to determine the whale's health, diet, sex, pregnancy status, microbiome, and details about its genetics. The AUVs will collect data regarding the whale and the underwater environment. They will use multi-hop underwater acoustic communication for delivering data to the vessel. By the time the boat arrives at the destination, the whale might not be at the location because of delays from the satellite image acquisition, NARW detection, and boat mobility. Thus, AUVs and drones would move in a determined pattern aimed at locating the whale detected before. Efficient algorithms for cooperative AUVs and drone path planning must be designed and implemented [11].

Some of the obtained data would be sent in real-time from the boat to the onshore monitoring center. The data would include the location information of the detected NARW and the location information of the AUVs and drones. The information regarding the whale will be used to inform ships and vessels in the area. Thus, they can reduce their speed or even divert them to reduce the chance of a fatal collision. The location information of the underwater vehicles and drones will be used for real-time routing aimed at tracking detected NARW. It will also be used to manage the Internet of aerial and underwater things, i.e., topology control and connectivity management. Moreover, the status of each drone and underwater vehicle (e.g., residual energy and energy usage) must be reported to the onshore monitoring center. This will be used to guarantee that drones and underwater vehicles will not have their battery depleted during a monitoring mission.

Finally, recorded multimedia data will be offloaded from the autonomous vessel for the onshore monitoring center. This transmission will be performed opportunistically, that is, when there is no ongoing real-time transmission. This will happen at the end of a monitoring mission. Non-real-time data transmission is performed at the end of the monitoring mission to avoid disturbing real-time data streams to be transmitted to the onshore monitoring center, as discussed above.

Current Challenges

The development of the envisioned CIoT system for whale monitoring faces many daunting challenges. The first challenge is the deployment of sufficient autonomous boats to cover a large area where NARWs migrate. There are COTS prototypes of autonomous sailboats that could be used (e.g., [6, 9]). One strategy would be the deployment of autonomous sailboats in strategy areas, such as over the Cabot Strait as illustrated in Fig. 4). The Cabot Strait is a major shipping route and is also used by NARWs when migrating from south areas to feeding grounds in the Gulf of St. Lawrence.

Another challenge is the design of docking stations for recharging drones on autonomous boats. Moore *et al.* [12] designed a physical docking-and-release station to recharge micro aerial vehicles. The charging station is deployed on a legged locomotion robot. The designed station has a landing area that triggers the docking and charging sequence when the micro aerial vehi-

cle lands. It has a 12V closed-cell, deep-cycle, lead-acid battery that provides acceptable performance under different temperature conditions and reduced potential of combustion. Once the vehicle lands, a set of claws is activated to prevent unwanted movements of the vehicle. This approach seems feasible to be installed in autonomous sailboats in the proposed CIoT system.

Data reliability is another critical challenge to be addressed in the envisioned CIoT system. Direct video streaming from underwater vehicles to autonomous boats is challenging due to the characteristics of the aquatic environment [13]. One solution can be the data collection and local storage at the underwater vehicle. Data will then be offloaded to the autonomous boat through a docking station (e.g., [14, 15]) when the vehicle returns to the boat.

Future Research Directions

This section discusses the future research directions on CIoT systems for whale monitoring.

- **Deployment:** In the envisioned architecture, the deployment of the considered devices will be fundamental. Internet of Underwater Things would need to be deployed at the locations where NARWs have commonly been encountered feeding. Thus, the proper deployment of the underwater sensor nodes must consider the need for multi-hop underwater communication for data delivery. The network topology must be pre-planned and deployed to guarantee adequate connectivity for reliable data transfer. Moreover, surface autonomous boats must be deployed at strategic locations with a high probability of finding an NARW. Such strategic locations can be determined from past aerial surveys and images obtained from satellites.
- **Localization and tracking:** Real-time localization and tracking are fundamental in the envisioned architecture. First, surface autonomous vessels must be able to determine their locations and report them in real-time for the onshore monitoring center. Their location information can be obtained through GPS receivers. Underwater vehicles must be capable of determining their locations when deployed by surface autonomous boats. They will use their location information for navigation when streaming video to the associated autonomous boat. Underwater vehicles' locations will be used at the onshore monitoring center for real-time monitoring and tracking. Furthermore, efficient localization algorithms must be implemented for the localization and tracking of AUVs. The AUVs' locations will be also used for navigation and real-time monitoring and tracking of their conditions.
- **Mobility management:** Mobility management will be fundamental to control the surface autonomous boats and the Internet of aerial and underwater things. Mobility management must be performed in a real-time manner. Upon a NARW is detected, the drones and underwater vehicles must move according to keep tracking the detected NARW. Thus, the trajectory of such devices will be determined by the movement of the monitored whale, which will be observed in real time through data collected by the nodes. Moreover, the mobility of drones and underwater must be performed in a way that will yet guarantee that they have connectivity with the surface autonomous vehicles that they are associated with.
- **Energy management:** Surface autonomous boats, drones, and underwater autonomous vehicles will operate autonomously when on mission at the ocean. Therefore, energy management is fundamental to guarantee that they can work properly. Solar panels and a battery recharging system must be implemented at each surface autonomous boat. Such a system must be used to recharge the boat, as well as the batteries of drones and underwater vehicles
- **Digital Twins for "things" management:** The envisioned CIoT system for NARW monitoring would rely on many autonomous entities, such as boats, drones, and underwater vehicles. Hence, it would be paramount the design and development of digital twins of those entities for the monitoring, fail detection, fail prediction, and remote maintenance of the computing systems of such entities.

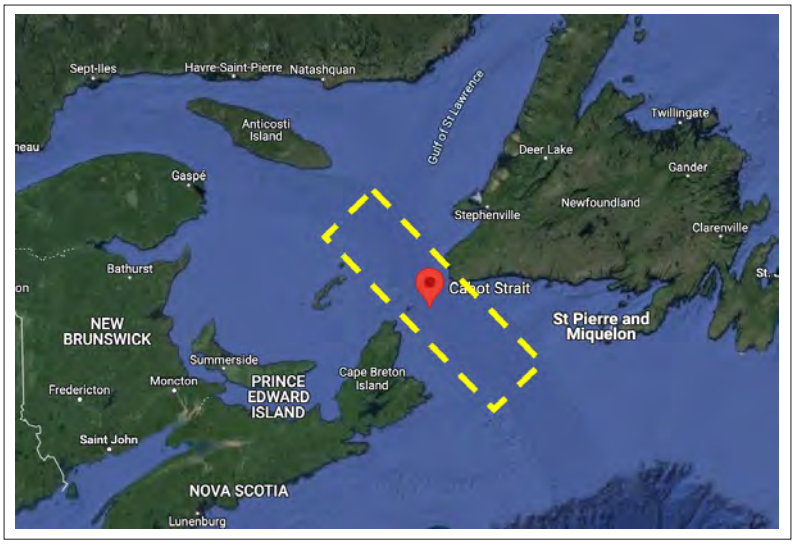


FIGURE 4. The Cabot Strait. Coordinates: 47° 15′ 00′′ *N* 59° 45′ 00′′ *W*.

Conclusion

The monitoring and tracking of North Atlantic Right Whales are fundamental for their preservation. To date, many efforts have been made aimed at the detection and tracking of NARWs. However, current solutions that rely on aerial and at-sea vessel surveys are limited. This article proposes a connected Internet of things architecture for the NARWs monitoring. The envisioned architecture is aimed at providing sensing and communication capabilities for detecting, monitoring, and tracking NARWs in remote and large geographic areas. We discussed the space, aerial, and underwater things to be involved in the envisioned architecture and the data flow and communication paradigms needed to implement the proposed CIoT architecture. Finally, we shed light on future research directions that must be addressed and aimed at developing building blocks of the envisioned architecture.

References

[1] I. F. Akyildiz and A. Kak, "The Internet of Space Things/Cubesats: A Ubiquitous Cyber-Physical System for the Connected World," *Computer Networks*, vol. 150, 2019, pp. 134–49.

[2] R. W. L. Coutinho and A. Boukerche, "North Atlantic Right Whales Preservation: A New Challenge for Internet of Underwater Things and Smart Ocean-Based Systems," *IEEE Instrumentation and Measurement Mag.*, 2021, pp. 1–17.

[3] M. Hodul *et al.*, "Individual North Atlantic Right Whales Identified from Space," *Marine Mammal Science*, vol. 39, no. 1, 2023, pp. 220–31.

[4] S. Kapoor, M. Kumar, and M. Kaushal, "Deep Learning Based Whale Detection from Satellite Imagery," *Sustainable Computing: Informatics and Systems*, vol. 38, 2023, pp. 100,858, pp. 1–11.

[5] A. Gaur *et al.*, "Whale Detection Enhancement Through Synthetic Satellite Images," *Proc. OCEANS 2023 — MTS/IEEE U.S. Gulf Coast*, 2023, pp. 1–7.

[6] A. P. F. Negreiros *et al.*, "Sustainable Solutions for Sea Monitoring with Robotic Sailboats: N-Boat and f-Boat Twins," *Frontiers in Robotics and AI*, vol. 9, 2022, pp. 1–12.

[7] B. Keller and T. Willke, "Snotbot: A whale of a Deep-Learning Project," *IEEE Spectrum*, vol. 56, no. 12, 2019, pp. 41–53.

[8] R. Y. Brogaard and E. Boukas, "Autonomous GPU-Based UAs for Inspection of Confined Spaces: Application to Marine Vessel Classification," *Robotics and Autonomous Systems*, vol. 172, 2024, p. 104590.

[9] A. P. D. de Araújo *et al.*, "General System Architecture and Cots Prototyping of an AIoT-Enabled Sailboat for Autonomous Aquatic Ecosystem Monitoring," *IEEE Internet of Things J.*, vol. 11, no. 3, 2024, pp. 3801–11.

[10] S. Herrería-Alonso *et al.*, "Improving Uplink Scalability of Lora-Based Direct-to-Satellite IoT Networks," *IEEE Internet of Things J.*, vol. 11, no. 7, 2024, pp. 12,526–35.

[11] Y. Wu *et al.*, "Cooperative Path Planning for Heterogeneous Unmanned Vehicles in A Search-and-Track Mission Aiming at an Underwater Target," *IEEE Trans. Vehic. Tech.*, vol. 69, no. 6, 2020, pp. 6782–87.

[12] B. Moore *et al.*, "Combined Docking-And-Recharging for A Flexible Aerial/Legged Marsupial Autonomous System," *Proc. 2023 IEEE Aerospace Conf.*, 2023, pp. 1–9.

[13] H. Luo *et al.*, "Underwater Real-Time Video Transmission via Wireless Optical Channels with Swarms of AUVs," *IEEE Trans. Vehic. Tech.*, vol. 72, no. 11, 2023, pp. 14,688–14,703.

[14] B. Pierre-Jean *et al.*, "Contactless Data Transfer for Autonomous Underwater Vehicle Docking Station," *Proc. the OCEANS 2021: San Diego — Porto*, 2021, pp. 1–5.

[15] J. Liu *et al.*, "A Review of Underwater Docking and Charging Technology for Autonomous Vehicles," *Ocean Engineering*, vol. 297, 2024, pp. 117,154, pp. 1–18.

Biographies

Rodolfo W. L. Coutinho (rodolfo.coutinho@concordia.ca) is an Associate Professor at the Department of Electrical and Computer Engineering at the Concordia University, Canada. He is the recipient of the the MSWiM Rising Star Award (2019), the Pierre Laberge Prize (2018) at the University of Ottawa, the Brazilian CAPES Best Thesis Award in the area of Computer Science (2018), the UFMG Best Exact and Earth Science and Engineering Thesis Award (2018), and the Best Doctoral Thesis Awards from the Thesis and Dissertation Competitions of the Brazilian Computer Society and the Brazilian Symposium on Computer Networks and Distributed Systems in 2018, respectively. He also received the Best Paper Award in IEEE GLOBECOM 2019 and IEEE MASCOTS 2014. He has served as TPC Co-Chair for IEEE and ACM international conferences, such as the IEEE GLOBECOM (IoT and Sensor Networks Symposium) 2024, IEEE ICC (AHSN Symposium) 2020, ACM MobiWac 2021 and 2020, and ACM DIVANET 2017. His research interests include Internet of Things, underwater networks, information-centric networking, pervasive computing, and mobile computing.

Azzedine Boukerche [FIEEE, FEiC, FCAE, FAAAS] (boukerch@site.uottawa.ca) is a Distinguished University Professor and holds a Canada Research Chair Tier-1 position with the University of Ottawa. He is the founding director of the PARADISE Research Laboratory and the DIVA Strategic Research Center, and NSERC-CREATE TRANSIT at the University of Ottawa. He has received the C. Gotlieb Computer Medal Award, Ontario Distinguished Researcher Award, Premier of Ontario Research Excellence Award, G. S. Glinski Award for Excellence in Research, IEEE Computer Society Golden Core Award, IEEE CS-Meritorious Award, IEEE TCPP Leaderships Award, IEEE ComSoc ComSoft and IEEE ComSoc ASHN Leaderships and Contribution Award, and the University of Ottawa Award for Excellence in Research. His current research interests include sustainable sensor networks, autonomous and connected vehicles, wireless networking and mobile computing, wireless multimedia, QoS service provisioning, performance evaluation and modeling of large-scale distributed and mobile systems, and large-scale distributed and parallel discrete event simulation.

IEEE Foundation

IEEE

Where technology and philanthropy intersect

Together, we deliver opportunity, innovation and impact across the globe.

As the philanthropic partner of IEEE, we translate the values of our members and donors into social impact. In collaboration with IEEE, we connect more than 200 member-led initiatives with financing, expertise and philanthropic guidance. Help advance the IEEE mission with a donation.

Funds and Programs:

- IEEE PES Scholarship Plus Initiative
- IEEE History Center and REACH
- EPICS in IEEE
- IEEE Smart Village
- And many more!

Join Us!

To find your program, visit
ieeefoundation.org/what-to-support

To make a donation, visit
ieeefoundation.org/donate

Energize

Engage

Educate

Illuminate

ieeefoundation.org

On the Support of the 2.4 GHz Band in the LoRaWAN Standard

Giampaolo Cuozzo, Riccardo Marini, Chiara Buratti, and Konstantin Mikhaylov

Abstract

The introduction of LoRa chipsets operating in the 2.4 GHz band paves the way to unprecedented performance enhancements compared to their sub GHz counterparts, attributed to factors such as the absence of duty cycle constraints and higher data rates. Despite its potential benefits for Internet of Thing (IoT) applications, the LoRa Alliance has not yet proposed the integration of this new frequency spectrum into the LoRaWAN standard. Addressing this gap, this article proposes a roadmap for the evolution of the LoRaWAN standard, outlining three stages for seamless integration of the 2.4 GHz LoRa version. These stages are sequenced based on implementation complexity, starting from the current LoRaWAN standard (Stage 0), moving to the coexistence of two separate LoRaWAN networks (Stage 1), and ending with a single LoRaWAN network capable of supporting both sub GHz and 2.4 GHz bands (Stage 2). Additionally, the document enumerates all possible implementation options for each stage and outlines the main modifications required in the documents of the LoRaWAN standard. Through LoRaWAN-compliant simulation results, we demonstrate the performance advantages of the proposed multi-band approach over the existing LoRaWAN standard for the first stage of the suggested roadmap. Finally, the article discusses the challenges associated with the proposed roadmap and identifies corresponding research gaps to be addressed in the future.

Introduction

LoRaWAN is the prominent Low Power Wide Area Network (LPWAN) technology and one of the most widely spread solutions for Internet of Things (IoT) applications [1], where sensors and/or actuator boards, referred to as End Devices (EDs), communicate with the Network Server (NS) by means of fixed radio stations, also known as Gateways (GWs) [2]. The success of LoRaWAN stems from its design, which utilizes LoRa technology at the Physical (PHY) layer [3] to minimize complexity, cost, and energy consumption while maximizing transmission range. The standardization of LoRaWAN was initiated by Semtech, an American company holding the patent for the synthesizer used to generate the modulated signal, and later adopted by the LoRa Alliance, an organization of companies that leads the standardization and harmonization processes of LoRaWAN.

LoRaWAN works in the sub GHz Industrial, Scientific and Medical (ISM) spectrum, where the bands are allocated on a country-by-country basis (Fig. 1). Despite its widespread adoption, LoRaWAN currently faces limitations for highly demanding IoT applications (e.g., smart manufacturing) due to the relatively narrow bands available in the sub GHz spectrum and duty cycle constraints. As a first step to overcome these issues, in 2017, Semtech introduced LoRa chipsets operating in the 2.4 GHz ISM spectrum [4]. However, at the time of writing, this frequency range is not included in the standard LoRaWAN frequency plans, despite its potential benefits given by the larger bandwidths, absence of duty cycle constraints, and improved miniaturization properties [5]. Furthermore, this technology might have the potential to outperform other state-of-the-art solutions operating in the same band, such as Wi-Fi and 802.15.4, thanks to its inherent flexibility and robustness against interference provided by chirp-based modulation [3].

This key contribution of this work is the roadmap for evolving the LoRaWAN standard to incorporate the 2.4 GHz LoRa version as an additional frequency range to be used. The proposed roadmap comprises three stages based on implementation complexity and integration simplicity. Stage 0 corresponds to the current LoRaWAN standard operating solely in the sub GHz ISM spectrum. In Stage 1, we envisage the deployment of 2.4 GHz GWs as hot-spot coverage extensions to address the shorter transmission range associated with this frequency band. Subsequently, Stage 2 leverages the concept of Relays (RLs) recently introduced by the LoRa Alliance [6], while enabling and extending it to 2.4 GHz band. This empowers EDs to operate in this frequency as forwarding entities to maximize network performance. Moreover, the article outlines all possible implementation options for each stage, spanning from architectural to protocol aspects, and identifies the main modifications required in the current LoRaWAN standard documents to support the proposed roadmap. Through LoRaWAN-compliant simulation results, and as a means to support our claims, we demonstrate the illustrative performance advantages of the proposed multi band approach over the existing LoRaWAN standard for Stage 1 of the roadmap. As a further notable contribution, the document finally discusses the challenges associated with the proposed roadmap and identifies research gaps for future exploration.

ISSN: 2576-3180
Digital Object Identifier: 10.1109/IOTM.001.2400054

Giampaolo Cuozzo and Riccardo Marini are with CNIT, Italy; Chiara Buratti is with the University of Bologna, Italy; Konstantin Mikhaylov is with the University of Oulu, Finland.

The structure of the article is as follows. First, we provide a brief introduction to the fundamental principles of LoRa and LoRaWAN. Building upon this foundation, we review the current literature concerning solutions for deploying LoRaWAN and LoRa at 2.4 GHz together. Subsequently, we detail the proposed roadmap, enumerate implementation options, and identify necessary modifications to the existing LoRaWAN standard. We then employ a standard-compliant simulator to demonstrate the potential benefits of integrating the 2.4 GHz band into the LoRaWAN standard. Finally, we list the research topics prompted by the proposed roadmap and draw conclusions.

Technical Background and Recent Developments

LoRa and LoRaWAN

In the LoRa modulation scheme, a set of M distinct chirp signals covers a designated frequency sweep interval, referred to in LoRa terminology as Bandwidth (BW). Specifically, the relationship $M = 2^{SF}$ holds, where Spreading Factor (SF) is an integer defining the temporal characteristics of each chirp. The larger the SF, the longer the transmission time, or Time on Air (ToA), while simultaneously boosting receiver sensitivity. LoRa also incorporates Forward Error Correction (FEC) techniques to balance reliability and transmission times. In precise terms, the LoRa frame, that is, the LoRa PHY Protocol Data Unit (PDU), comprises, in addition to the Cyclic Redundancy Check (CRC) bits, a preamble (for detection and synchronization), an optional header, and the payload (i.e., the useful data).

Coming to the upper layers, the LoRaWAN protocol is based on ALOHA, thus an ED randomly selects one of the available frequency channels and sends a LoRa frame whenever it has new data ready. Rather than being associated with a specific GW, EDs are linked to a NS, hence their transmissions are received by all the reachable GWs. The NS is responsible for generating Acknowledgments (ACKs), eliminating potential duplicates, managing join requests, and overseeing and optimising the overall network. The two additional components are Application Servers (ASs) and Join Server (JS). The ASs fetches the decoded application data from the NS, and the JS manages authentication.

As far as the LoRaWAN network architecture is concerned, EDs can also work as RLs and act as network extenders [6], improving even more the coverage of LoRaWAN networks. RLs operate by receiving messages from EDs and subsequently retransmitting them to a GW. This functionality enables far EDs to communicate with the GW, overcoming distance limitations.

LoRaWAN is recognized for its sub GHz versions, operating in license free ISM bands, such as EU868 and US915 for Europe and the US, respectively [7]. As shown in Fig. 1, these bands are not harmonized around the globe, which complicates production and hampers roaming. There are also regional specific limitations that affect the LoRaWAN performance, for example, the EU868 MHz spectrum (the sub GHz reference spectrum in this work) mandates the implementation of three default channels in each ED. Additionally, within this spectrum, regulatory constraints impose a duty cycle limitation of 1% for the default channels to mitigate potential interference caused by devices.

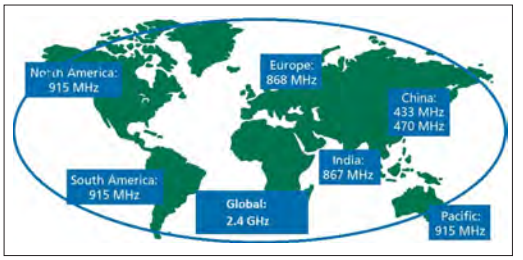


FIGURE 1. Comparison of LoRa bands across the world according to the regional parameters specification [7] and 2.4 GHz ISM band.

On the other hand, the 2.4 GHz band is available almost everywhere around the globe and features fewer restrictions on its use than sub GHz band which are compatible with the LoRaWAN working principle, thereby motivating the contribution contained in this article.

State of the Art

There exist a few works in the literature, mostly experimental, which envision the introduction of the 2.4 GHz PHY layer and its cooperation with already existing sub GHz bands into the standard. In [8], an experimental comparison between LoRa sub GHz and LoRa 2.4 GHz in both indoor and outdoor scenarios is carried out, and a discussion on the research challenges that need to be addressed to fully take advantage of this new technology is given. In [9], an experimental setup evaluating the LoRaWAN protocol stack on top of the LoRa 2.4 GHz is analyzed and a new Time Slotted Channel Hopping protocol is proposed. Authors in [10] exploit the well known ns-3 simulator to analyze network performance in different frequency bands. However, none of the above cited works discuss how integration of the 2.4 GHz in the specification can be done. This is the gap we focus on in the following.

Multi-Band LoRaWAN

Motivation

The introduction of a multi band LoRaWAN technology, that is, a LoRaWAN network that can work both in the sub GHz and 2.4 GHz band, presents an exciting opportunity to elevate mobile radio networks within the IoT landscape. This innovation heralds new horizons for enhancing network performance and augmenting deployment flexibility, accounting for the evolving needs of IoT applications. Thanks to its characteristics [5], LoRa 2.4 GHz is particularly well suited for applications with stringent traffic requirements (e.g., real time data collection in industry plants). Additionally, it allows for the reduction in antenna size, further enhancing deployment flexibility, along with an improved energy efficiency due to transmit power limitations and smaller transmission times. Remarkably, dual band solutions also help improve the network scalability and better tailor to propagation conditions and application requirements. Given these compelling advantages, it is worth investigating approaches that facilitate the seamless coexistence of both the regulated sub GHz version and the new 2.4 GHz iteration within a unified LoRaWAN standard. By leverag-

Coming to the upper layers, the LoRaWAN protocol is based on ALOHA, thus an ED randomly selects one of the available frequency channels and sends a LoRa frame whenever it has new data ready.

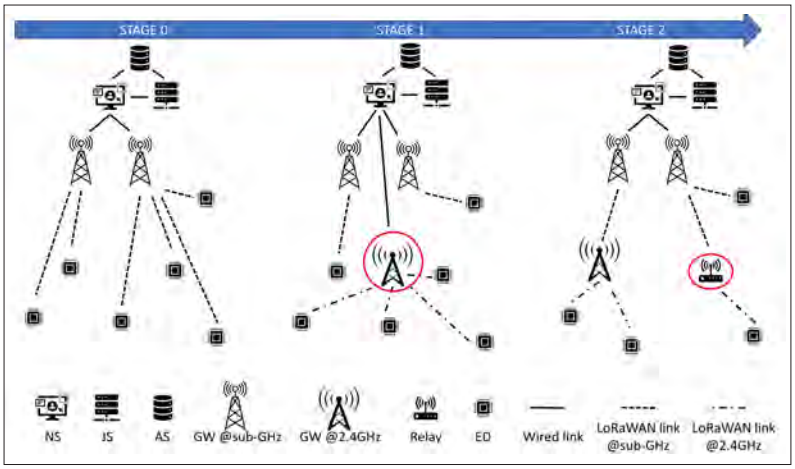


FIGURE 2. Suggested roadmap of the LoRaWAN standard to enable support of the 2.4 GHz LoRa version. Multiple links between EDs and GWs are not considered (even if possible) for simplifying the figure, and only the shortest ED-GW link is shown. Red circles identify the main additional components of Stages 1 and 2 in comparison to the state of the art Stage 0.

ing the strengths of this technology, LoRaWAN stands poised to revolutionize IoT connectivity and drive innovation in various industries, leading to a general increase in network performance.

Roadmap

Figure 2 depicts our vision of how the LoRaWAN standard should evolve to support the 2.4 GHz LoRa version. In particular, we envision three main stages:

- **Stage 0:** This stage aligns with the existing LoRaWAN standard [7], wherein EDs communicate with GWs using LoRa technology in the sub GHz band. The GWs then conveys this information through wired connections to and from the NS (it is worth noting that, to maintain simplicity throughout the text, we will refrain from detailing the various components of the NS, including the home, serving, and forwarding NS). The NS, in turn, can be linked to the JS and AS for managing security, authentication, and other relevant aspects;
- **Stage 1:** This is the first evolution of Stage 0, where GWs and EDs can also embed 2.4 GHz radio modules to properly exploit the benefits of this band.
- **Stage 2:** This is the final stage, where the concept of RL at 2.4 GHz is further introduced, to also forward LoRa frames in this band.

Implementation Options

In light of the macroscopic view of the proposed roadmap, we hereby present the different implementation options for Stages 1 and 2 only, as Stage 0 represents the current LoRaWAN standard.

Stage 1: The addition of the 2.4 GHz radio module in both, EDs and GWs, can be done in different ways. We enumerate the implementation options (which can also be considered as development steps) in ascending order of difficulty of integration:

- *Option 1.1:* Both GWs and EDs communicate via LoRaWAN at 2.4 GHz. Remarkably, this is not foreseen by the current LoRaWAN specification, as only the PHY layer at sub GHz is envisioned;
- *Option 1.2:* There are two types of GWs and EDs, one working at sub GHz and the other at 2.4 GHz, so that there exist separate GW-ED links that can be managed either by a single or multiple NSs.
- *Option 1.3:* There are two types of EDs, one working at sub GHz and the other at 2.4 GHz, whereas GWs are equipped with both radio modules to provide connectivity to all categories of EDs;
- *Option 1.4:* There are two types of GWs, one working at sub GHz and the other at 2.4 GHz, whereas EDs are equipped with both radio modules and properly select (or are forced to choose) the frequency spectra to be used. This is the version illustrated in Fig. 2;
- *Option 1.5:* Both GWs and EDs are equipped with both radio modules, that is, sub GHz and 2.4 GHz, and some logic at the NS determines which one to use;

Stage 2: The concept of RLs in the current LoRaWAN standard [6] requires that EDs, RLs, and GWs always operate in the same frequency spectrum. Differently, in this article, we foresee that GWs operate in the sub GHz band, whereas there are different possible options for RLs and EDs:

- *Option 2.1:* RLs are equipped with 2.4 GHz and sub GHz radios, whereas there are two types of EDs, each supporting only one frequency band. EDs working at 2.4 GHz communicate with RLs that, in turn, communicate with GWs via LoRaWAN at sub GHz. The EDs at sub GHz communicate with the GW directly.
- *Option 2.2:* Same as option 2.1, but RLs also forward data from/to EDs working in the sub GHz band. This option sets overhead on the sub GHz band, as it should account for two radio links, namely RLs-GWs and EDs-RLs;
- *Option 2.3:* RLs and EDs are equipped with radio modules for both frequencies so that all possibilities foreseen by the previous options are feasible.

Standard Modifications

In this following, we highlight the essential modifications needed for the LoRaWAN standard to align with the roadmap proposed in this article.

Specifically, Table 1 offers a summary of the primary modifications to the chapters of the current LoRaWAN standard required to implement Stages 1 and 2 for the various implementation options previously outlined. Obviously, the LoRaWAN regional parameters [7] should be adjusted to introduce the 2.4 GHz band, by specifying both the possible range of values (e.g., channel frequencies, data rates, transmit powers, maximum payload size, receive windows, etc.) and the corresponding PHY layer characteristics [4]. However, for both stages, the major changes in [11] are required, as the responsibilities of EDs, RLs, GWs, and NS should be extended to elucidate who determines the utilization of dual radios, the methodology employed for such decisions, and the collection of data to support them. Moreover, the modified standard should also describe the role of the multi bands in association and roaming procedures. In particular, from the network architecture viewpoint and considering Stage 2, the locations of EDs acting as RLs can either:

- Be chosen a priori to provide sufficiently high performance, that is, statically creating links between EDs, RLs and GWs

- A logic might be implemented in one of the network elements to dynamically transform EDs into RLs and create EDs-RLs links whenever it is needed. Additionally, the ED activation procedure, both Over-the-Air-Activation (OTAA) and Activation By Personalization (ABP) mode, should be properly modified as the NS should also acquire the knowledge on the bands supported by EDs, despite the rationale of such procedures can be seamlessly transitioned to the 2.4 GHz spectrum.

Another important modification of [11] refers to the criteria for communication mapping. Indeed, data communication may be allocated to either one frequency band or both (e.g., to increase reliability at the expense of additional complexity in managing duplicates), as well as it is possible to foresee a possible frequency split between uplink and downlink. Finally, the L2 and Relay specifications (i.e., [2] and [6]), should account for implementation details such as the introduction of new Medium Access Control (MAC) commands to switch between the two bands, or to keep the NS informed of the bands currently used by EDs.

Selected Numerical Results

In the following, some exemplary numerical results will be discussed to highlight how the proposed roadmap can improve network performance compared with the current LoRaWAN standard (i.e., Stage 0). For the sake of brevity, we will consider only Stage 1, Options 1.1 and 1.2, but the discussion can be extended to the other cases as well.

Simulations have been realized exploiting the open source LoRaWANSim simulator described in [12], properly modified to match the context of this work.

System Model and Simulation Setup

Scenario: We consider a scenario where a GW is located at the center of a square area with side length of A and N EDs are randomly and uniformly distributed around. The simulation lasts T seconds.

Channel Model: We assume the 3rd Generation Partnership Project (3GPP) Urban Macro (UMa) channel model described in TR 38.901 [13], which is valid for frequencies from 0.5 to 100 GHz.

Traffic Model: We focus on uplink traffic, which we assume to be periodic with period T_U and a payload of B bytes.

End Device Configuration: EDs work in *class A* mode [2]; hence, following an uplink transmission, they initiate two downlink receive slots, labelled Receive Window 1 (RX1) and Receive Window 2 (RX2), after specific fixed intervals of 1 second and 2 seconds (as per standard practice). For a more fair comparison, we assume that EDs exploit just three channels in both bands (specifically, we assume sub GHz EDs work in the EU868 MHz band). EDs employ the Adaptive Data Rate (ADR) algorithm [14], designed to dynamically configure the SF and transmit power used by EDs during their transmissions. We assume that the entire set of SFs is available at the ED side, which is 7–12 for sub GHz EDs and 5-12 for 2.4 GHz EDs, respectively. Finally, they operate in *unconfirmed mode*, so their transmissions do not require an ACK sent by the NS via GWs.

	LoRaWAN documentation	Options	Chapters
Stage 1	Regional Parameters [7]	1.1–1.5	2, 4
	Main Specifications [2]	1.2, 1.4–1.5 1.2-1.5	5 6
	Backend Specification [11]	1.1–1.5	3, 6–9, 11, 18
Stage 2	Regional Parameters [7]	2.1–2.3	3, 4.4
	Backend Specification [11]	2.1–2.3	3, 6–9, 18.3
	Relay Specification [6]	2.1–2.3	3, 8, 10

TABLE 1. Main modifications to the chapters of the current LoRaWAN standard to implement Stages 1 and 2 of the proposed multi band roadmap as a function of the chosen implementation option.

Simulator Working Principle: In the simulation, an uplink packet is considered correctly received, taking into account two impairments: noise and interference. Regarding the former, we model the entire transmitter-receiver chain. Therefore, based on the PHY layer parameters and the Signal to Noise Ratio (SNR) characterizing a specific link, the PHY layer part of the simulator determines whether the currently transmitted packet is received accurately. As for the interference, we take into account the possibility that colliding LoRa frames may still be correctly received if the corresponding Signal to Interference Ratio (SIR) exceeds a given threshold (which changes according to the SFs used for the transmissions), hence accommodating for the capture effect. This model has been largely used in the literature, as described in detail in Section V of [12].

Numerical Results

Network performance is evaluated in terms of success probability and network throughput as a function of the offered traffic, O [bit/s], defined as

$$O = \frac{B \cdot N}{T_U}.$$

The success probability, P_S, represents the percentage of frames correctly received by the GW and it is the ratio between the number of correctly received frames over the number of transmitted ones. The network throughput, S, has been defined as the number of data bits correctly received by the GW from all EDs throughout the simulation, measured in bit/s.

Parameters used during simulation are reported in Table 2. The selected results are depicted in Fig. 3 and Fig. 4. For each figure, three possibilities are presented: Stage 0 (i.e., the current LoRaWAN standard); Stage 1, Option 1.1; Stage 1, Option 1.2.

Figure 3 shows the success rate, P_S, as a function of the offered traffic, O. As one can notice, as the offered traffic increases, the success probability decreases. The reason behind such behavior is twofold: on one hand, increasing the number of EDs, N, will impact the number of collisions during frame transmissions in the network; on the other hand, with a larger B, the ToA of the frame will increase, leading to a longer vulnerability interval, hence more collisions. Moreover, the reduced data rate and the duty cycle constraints of LoRaWAN at EU868 MHz (Stage 0) produce even longer ToA, resulting in more collisions and fewer transmissions for the same amount of

Despite the inclusion of an additional transceiver is a consolidated procedure that does not significantly increase the overall price of the device, justifying this effort requires substantial performance improvements.

Parameter	Value
Area side length (A)	1 km
LoRa Frame Periodicity (T_U)	10 s
LoRa Frame Size (B)	10 B
Simulation duration (T)	120 s
Frequency Sweep Interval (BW)	125 (@EU868 MHz) kHz 203 (@2.4 GHz) kHz

TABLE 2. Simulation parameters.

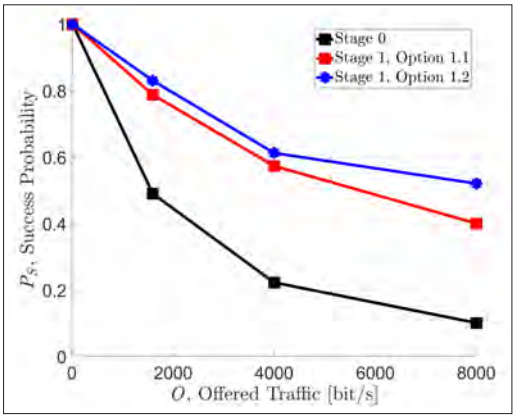


FIGURE 3. Success probability as a function of the offered traffic, O.

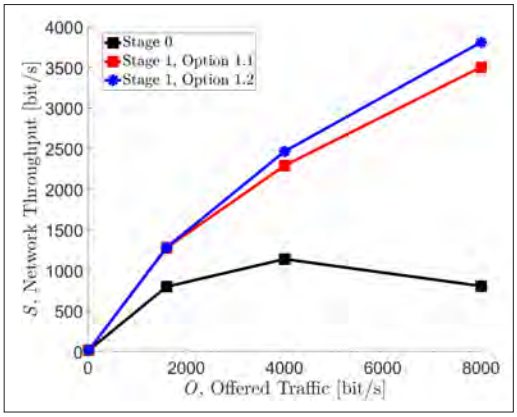


FIGURE 4. Network throughput, S, as a function of the offered traffic, O.

time w.r.t. the other cases, making the standard solution the worst possible choice. On the other hand, introducing the 2.4 GHz band (Option 1.1), the success rate increases thanks to the reduced ToA. The joint solution (Option 1.2) results in even better performance since sub GHz and 2.4 GHz EDs' transmissions do not collide, reducing the overall number of collisions in the network.

The success probability has a direct impact on the network throughput, as shown in Fig. 4. As it can be seen, Stage 0 shows the usual ALOHA behavior, with a peak around O = 4000 bit/s. Again, moving towards Stage 1, the network performance increases and the peak shifts further right, showing the benefit of the 2.4 GHz band introduction. As can be seen, for O = 8000 bit/s the suggested Stage 1 modifications enable supporting 3.5 higher overall network throughput.

Open Challenges and Research Directions

The proposed roadmap presents a promising avenue for mobile radio networks. However, it also introduces multiple challenges for design and optimizing the operation both from a network perspective (e.g., the development of dual band ADRs) and from the ED point of view (e.g., determining the starting band to use for OTAA). These challenges underscore the necessity to explore new avenues of research. In the following, we will outline the primary research directions that the proposed roadmap unveils.

Optimal Configurations for 2.4 GHz LoRaWAN: One of the first challenges to be solved is the optimal selection of communication configurations to be used for LoRaWAN 2.4 GHz operation, that is, determining the number and frequency of channels, specification of the bandwidth(s), SFs, and transmit power levels, the number of RX windows and respective delays, etc. Depending on the application and the level of integration with sub GHz LoRaWAN, there can even be several different configuration options.

Interference Mitigation: One further significant challenge lies in potential interference from existing networks operating within the 2.4 GHz band, such as Wi-Fi or Bluetooth. The coexistence of these networks is likely to happen in different scenarios (e.g., smart homes), impacting the performance and reliability of the proposed technology. In this regard, there already exist studies (e.g., [15]), that provide indications of the resilience of LoRa at 2.4 GHz from co-channel interference. However, further work may be directed toward implementing advanced interference mitigation strategies at the network and MAC levels.

Cost/Performance Tradeoffs: The proposed approach may incur higher costs due to the addition of radio modules operating in the 2.4 GHz band at EDs and/or GWs. Despite the inclusion of an additional transceiver is a consolidated procedure that does not significantly increase the overall price of the device, justifying this effort requires substantial performance improvements. Without tangible benefits, market viability may be compromised. Notably, our numerical findings demonstrate a remarkable improvement in network throughput when comparing Stage 1, option 1.2 with the state of the art LoRaWAN technology (refer to Fig. 4). Nevertheless, future studies, potentially including experimental trials, are needed to further confirm these preliminary observations and to bolster the proposed roadmap for the LoRaWAN standard.

Dynamic Allocation of End Devices to Relays: Another crucial research direction involves the dynamic allocation and handovers of EDs to RLs based on different factors, such as, traffic load, channel conditions, and proximity to RLs. Adaptive allocation strategies, including multi band adaptive data rate (ADR) can enhance network efficiency and resource utilization, ensuring optimal performance under diverse operating conditions and by mitigating the downsides of both LoRaWAN and LoRa at 2.4 GHz.

MAC Protocol Design: Designing more efficient MAC protocols than the approach similar to ALOHA of LoRaWAN for communication between EDs and RLs at 2.4 GHz (e.g., with specific application needs in mind), presents an intriguing research

opportunity. Despite potential deviation from standard compliance, such protocols must offer substantial advantages to justify their adoption.

Multi-Hop Design Complexity: Designing the network architecture becomes intricate when considering wide scenarios requiring more than a single hop (i.e., more than one RL), to connect EDs with GWs. Managing multiple hops while maintaining sufficiently high performance levels (e.g., in terms of energy efficiency, reliability, and network throughput) poses a significant and unprecedented challenge in the routing strategies of LoRaWAN networks. We advocate for the comparison of LoRaWAN architectures incorporating multiple hops with, at minimum, the performance offered by the Stage 0 version. This comparison is essential to ascertain any potential advantage associated with the proposed approach. Nevertheless, even when considering a single level of RLs, the development of efficient RL selection algorithms is mandatory. These procedures should intelligently and dynamically identify RLs based on factors such as signal strength, interference levels, and network topology to optimize performance and reliability.

Security: The introduction of dual-band devices imposes some security concerns. First, it becomes mandatory to ensure transparent sequence numbering across interfaces for dual-mode EDs to mitigate replay attacks in alternate bands. Additionally, the introduction of additional network elements, i.e., GWs at 2.4 GHz and RLs, extends the number of devices that are physically accessible from malicious users. Hence, enhancing defenses against denial-of-service attacks and safeguarding against physical tampering or hijacking of equipment is imperative.

Conclusions

This article has reported how the potential integration of LoRa chipsets operating at 2.4 GHz into the existing LoRaWAN standard can be enabled. By proposing a roadmap consisting of three distinct stages, starting from the current LoRaWAN standard (Stage 0), moving to the coexistence of two separate LoRaWAN networks (Stage 1), and ending with a single LoRaWAN network capable of supporting both sub GHz and 2.4 GHz bands (Stage 2), we have elucidated the feasibility and benefits of this proposed approach.

Specifically, for Stages 1 and 2, we discussed the different implementation options and necessary modifications to Stage 0 documentation, spanning from architectural to protocol aspects, with the objective of providing practical insights for stakeholders involved in the standardization process. By leveraging LoRaWAN compliant simulation results, we have empirically demonstrated the performance advantages of the proposed multi band approach, particularly in Stage 1 of the roadmap, where GWs equipped with radio modules at 2.4 GHz can be deployed to boost success probability and network throughput.

Looking ahead, our work also highlights several research directions and open questions that warrant further investigation for the research community. These include optimal selection of configurations for 2.4 GHz LoRaWAN, interference mitigation strategies, cost/performance tradeoffs, dynamic allocation of EDs to RLs, MAC protocol design, and multi hop network architecture. As the IoT landscape continues to evolve, addressing these challenges and opportunities will be essential for unlocking the full potential of LoRaWAN technology.

References

[1] A. Ikpehai *et al.*, "Low-Power Wide Area Network Technologies for Internet-of-Things: A Comparative Review," *IEEE Internet of Things J.*, vol. 6, no. 2, 2018, pp. 2225–40.
[2] LoRa Alliance, LoRaWAN® L2 1.0.4 Specification (TS001-1.0.4), 2020.
[3] G. Pasolini, "On the LoRa Chirp Spread Spectrum Modulation: Signal Properties and Their Impact on Transmitter and Receiver Architectures," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, 2021, pp. 357–69.
[4] S. Corporation, "SX1280/1281 Data Sheet DS.SX1280-1.W.APP," 2018.
[5] R. Marini and G. Cuozzo, "A Comparative Performance Analysis of LoRaWAN in Two Frequency Spectra: EU868 MHz and 2.4 GHz," *2023 Joint European Conf. Networks and Commun. & 6G Summit (EuCNC/6G Summit)*, 2023, pp. 1–6.
[6] LoRa Alliance, LoRaWAN® Relay Specification TS011-1.0.0, 2022.
[7] —, LoRaWAN® Regional Parameters RP002-1.0.4, 2022.
[8] G. H. Derevianckine *et al.*, "Opportunities and Challenges of LoRa 2.4 GHz," *IEEE Commun. Mag.*, 2023, pp. 1–7.
[9] M. Schappacher, A. Dant, and A. Sikora, "Implementation and Validation of LoRa-Based Systems in the 2.4 GHz Band," *2021 IEEE 4th Int'l. Conf. Advanced Information and Commun. Technologies*, 2021, pp. 106–11.
[10] Masek, Pavel *et al.*, "Performance Analysis of Different LoRaWAN Frequency Bands for mMTC Scenarios," *2022 45th Int'l. Conf. Telecommun. and Signal Proc.*, 2022, pp. 417–20.
[11] LoRa Alliance, LoRaWAN® Backend Interfaces Technical Specification (TS002-1.1.0), 2020.
[12] R. Marini *et al.*, "LoRaWANSim: A Flexible Simulator for LoRaWAN Networks," *Sensors*, vol. 21, no. 3, 2021.
[13] 3GPP, "Study on Channel Model for Frequencies from 0.5 to 100 GHz (Release 16)," TS 38.901, 2019.
[14] R. Marini, W. Cerroni, and C. Buratti, "A Novel Collision-Aware Adaptive Data Rate Algorithm for LoRaWAN Networks," *IEEE Internet of Things J.*, vol. 8, no. 4, 2021, pp. 2670–80.
[15] G. Cuozzo, C. Buratti, and R. Verdone, "A 2.4-GHz LoRa-Based Protocol for Communication and Energy Harvesting on Industry Machines," *IEEE Internet of Things J.*, vol. 9, no. 10, 2021, pp. 7853–65.

Biographies

Giampaolo Cuozzo [M'23] (giampaolo.cuozzo@wilab.cnit.it) received from the University of Bologna a B.S with honors in electronics and telecommunications engineering in 2017, an M.S with honors in telecommunications engineering in 2019, and a Ph.D. degree in electronics, telecommunications, and information technologies engineering (ET-IT) in 2023. He is currently Head of Research at the National Laboratory of Wireless Communications (WiLab) of CNIT (the National, Inter-University Consortium for Telecommunications). His research activity is focused on the study, development, and validation of 6G networks for the Industrial Internet of Things.

Riccardo Marini [M'23] (riccardo.marini@wilab.cnit.it) received the M.Sc with honors in Telecommunications Engineering and got the Ph.D. in Electronics, Telecommunications and Information Technologies Engineering (ET-IT) at the University of Bologna in 2019 and 2023, respectively. He is currently Head of Research at the National Laboratory of Wireless Communications (WiLab) of CNIT (the National Inter-University Consortium for Telecommunications). His research activities focus on urban and rural Internet of Things, with emphasis on LPWAN technologies and UAV networks, as well as Machine Learning techniques for wireless networks.

Chiara Buratti [M'06] (c.buratti@unibo.it) received the Ph.D. in Electronics, Information Technologies and Telecommunications Engineering at the University of Bologna in 2009. She is Associate Professor at University of Bologna. Her research interests are on wireless sensor networks and IoT, with emphasis on MAC and routing protocols and on 3D networks. She co-authored approx. 120 papers. She has been PI of the COST Innovators Grant, Immunet. She was the main proponent of the Cost Action CA20120, INTERACT, and she is currently the Vice-Chair and Grant Holder of the Action.

Konstantin Mikhaylov [S'10, M'18, SM'19] (konstantin.mikhaylov@oulu.fi) received Dr.Sc. (Eng) in Telecommunication from the University of Oulu, Finland, in 2018. Currently he is an Assistant Professor (Tenure Track) for Convergent IoT Communications for Vertical Systems with the Centre for Wireless Communications at the University of Oulu. The major focus of his research is on the radio access and beyond-access technologies for massive and dependable IoT, and the matters related to the design and use of IoT systems. He has authored and co-authored over one hundred research papers on wireless connectivity for IoT, IoT devices and systems design, and applications.

Open Access funding provided by University of Bologna within the CRUI CARE agreement.

Opportunities for Intelligent Reflecting Surfaces in 6G Empowered V2X Communications

Wali Ullah Khan, Asad Mahmood, Arash Bozorgchenani, Muhammad Ali Jamshed, Ali Ranjha, Eva Lagunas, Haris Pervaiz, Symeon Chatzinotas, Björn Ottersten, and Petar Popovski

Abstract

The applications of upcoming sixth-generation (6G) empowered vehicle-to-everything (V2X) communications depend heavily on large-scale data exchange with high throughput and ultra-low latency to ensure system reliability and passenger safety. However, in urban and remote areas, signals can be easily blocked by various objects. Moreover, the propagation of signals with ultra-high frequencies, such as millimetre waves and terahertz communications, is severely affected by obstacles. To address these issues, the Intelligent Reflecting Surface (IRS), which consists of nearly passive elements, has gained popularity because of its ability to intelligently reconfigure signal propagation in an energy-efficient manner. Due to the promise of ease of deployment and low cost, IRS has been widely acknowledged as a key technology for both terrestrial and non-terrestrial networks to improve wireless coverage signal strength, physical layer security, positioning accuracy, and reduce latency. This article first describes the introduction of 6G empowered V2X communications and IRS technology. Then, it discusses different use case scenarios of IRS enabled V2X communications and reports recent advances in the existing literature. Next, it focuses our attention on the scenario of vehicular edge computing involving IRS enabled drone communications in order to reduce vehicle computational time via optimal computational and communications resource allocation. Finally, this article highlights current challenges and discusses future perspectives of IRS enabled V2X communications in order to improve current work and spark new ideas.

Introduction

The sixth-generation (6G) empowered vehicle to everything (V2X) communications is essential to smart city transportation systems. Robust wireless connections and cutting-edge sensors will completely transform the safety and comfort of the existing transportation systems [1]. The future transportation industry will incorporate a wide range of technologies, including those for passenger and driver protection, autonomous driving, traffic management, and passenger amusement. By providing pervasive connectivity, secure data sharing, energy-efficient transmissions, and quick computation, 6G wireless technology is the backbone of the transportation industry. Furthermore, the 6G transportation system will offer terabit-per-second data rates, which are exceptionally high. As a result, the latency of wireless communications can be reduced to under 1 millisecond, and the packet delivery ratio can be increased to ≈ 100% [2]. 6G will be enabled by technologies including intelligent reconfigurable surfaces (IRS), terahertz (THz) communications, blockchain, ambient backscatter communications, and artificial intelligence.

Besides the promise of the above features, V2X communications also face several challenges. For example, shadowing effects can significantly impact the efficiency and effectiveness of V2X communications due to obstacles like buildings in urban settings or hills and trees in rural areas. Therefore, limited energy reservoirs and spectrum resources would be the main challenges for large-scale V2X communications in 6G. Future V2X communications may also suffer from low transmission latency, unreliable wireless connectivity, and/or limited coverage. Moreover, high-velocity vehicles impact channel stability, having a negative impact on data rates. Accordingly, changing the position of drones in the air complicates communications even further. Keeping a high degree of energy efficiency in V2X communications while attempting to control the propagation and fading of THz signals is an open question. Driving safety and communications security are compromised by V2X communications that are unstable. It is essential to increase the range of communications and strengthen it in a sustainable manner.

The IRS has been seen as a potentially game-changing technology in 6G, with the ability to manipulate signal propagation and develop an intelligent radio environment [3]. Using reflection and programming, IRSs can alter the phase of incoming electromagnetic (EM) waves, allowing for the redesign of channels. IRS reflection can create a new propagation path around an obstacle that is impeding the direct Line-of-Sight (LOS) link between the source and destination. In conventional communications systems, re-engineering the transceiver is the only option for boosting system performance. The IRS adds a new design

Wali Ullah Khan, Asad Mahmood, Arash Bozorgchenani, Muhammad Ali Jamshed, Ali Ranjha, Eva Lagunas, Haris Pervaiz, Symeon Chatzinotas, Björn Ottersten, and Petar Popovski

Digital Object Identifier: 10.1109/IOTM.001.2300096

2576-3180/24/$25.00 © 2024 IEEE

parameter to wireless networks. Therefore, IRS technology can be used to enhance V2X communications and offer indirect LOS links that are both cost-effective and energy-efficient. Because significant performance gains are achieved only when the transceiver is close to the IRS, a permanently deployed IRS will limit its potential. Given the transient nature of vehicles, mobile IRS is viewed as a viable option for V2X networks.

Great potential exists for 6G empowered V2X communications thanks to IRS's ability to enable beyond LOS and energy-efficient communications. The IRS promises to help vehicle to infrastructure (V2I), vehicle to vehicle (V2V), vehicle to drone (V2D), and vehicle to satellite (V2S) communications in 6G networks by improving multipath propagation and expanding transmission coverage in high-frequency bands, i.e., millimeter wave (mmWave) and THz. The IRS is also simple to deploy due to its two-dimensional plane surface structure. Furthermore, the passive reflection mechanism enables IRS to operate in a low-energy-consumption mode, meeting the green 6G empowered V2X communications requirements. Recent hardware and material research indicate that it can control the reflection dynamically, allowing the IRS to perform real-time beamforming and serve multiple vehicles. IRS's ability to use reconfigurable passive beamforming to strengthen physical layer security in vehicular communications on the ground and in the air is a major advantage.

This article describes the IRS opportunities in 6G empowered V2X communications and highlights some existing problems for ground and aerial/space V2X communications. First, it discusses various use case scenarios of IRS enabled V2X communications and provide recent advances in the literature. Then, it presents a case study on vehicular edge computing (VEC) network involving IRS enabled cooperative drone communications, with the goal of reducing vehicle computational time using a new optimization framework. Before IRS can be widely used in 6G empowered V2X communications, some issues need to be resolved. We underline the challenges so as to provide direction for the implementation of IRS in terrestrial and non-terrestrial V2X communications. The rest of this article is structured as follows: a section contains use case scenarios as well as recent advances. We introduce a new optimization framework for minimizing computational time in VEC networks involving IRS enabled drone communications. We discuss unresolved issues and potential future research directions. We conclude with closing remarks.

IRS-Enabled V2X Communications: Use Case Scenarios and Recent Advances

In this section, we first highlight and discuss potential use case scenarios in 6G IRS enabled V2X communications.[1] Then we report and compare recent advances in IRS enabled V2X communications.

Use Case Scenarios of IRS-Enabled V2X Communications

The potential use case scenarios of V2X communications involving IRS are shown in Fig. 1, including V2I, V2V, V2D and V2S, respectively. In the following, we discuss these case scenarios in more detail.

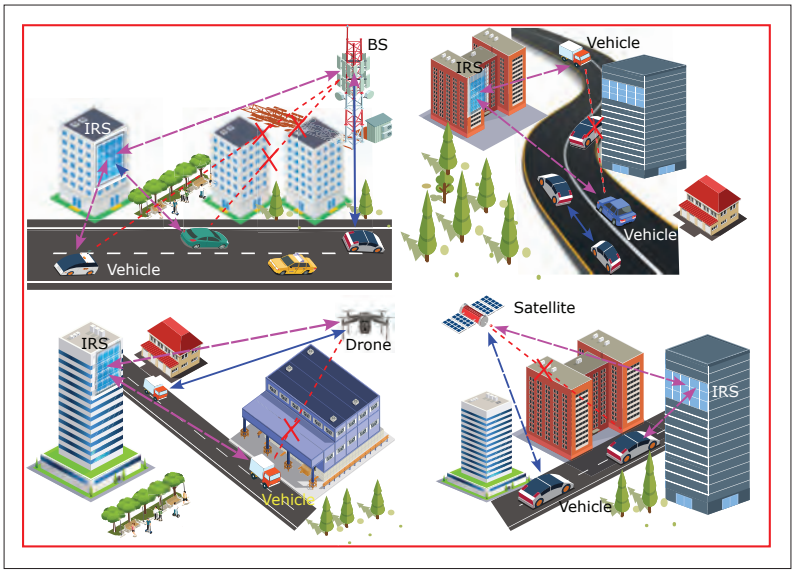


FIGURE 1. Potential use case scenarios of IRS enabled V2X communications, i.e., IRS enabled V2I communications, IRS enabled V2V communications, IRS enabled V2D communications, and IRS enabled V2S communications.

IRS Enabled V2I Communications: V2I communications can face the challenges of signal blockage and large-scale fading in urban areas. One of the traditional methods in such a use case scenario is to deploy relay devices to improve the received signal strength [1]. However, it requires extra power consumption. The IRS can intelligently reconfigure the signal toward the receiver, which extends wireless coverage and enhances energy efficiency in non-line-of-sight (NLOS) scenarios. In Fig. 1, we can see V2I communications where a base station (BS) communicates with multiple vehicles in an urban area, facing signal blockage due to high buildings, compromising their performance. In such a scenario, the IRS can be mounted strategically to assist the signal delivery and thus improve the system performance.

IRS Enabled V2V Communications: In the V2V use case scenario, the communications between different vehicles can be blocked by other vehicles/objects on the road and roadside [2]. Figure 2 illustrates an example of a V2V scenario, where the transmissions between two vehicles are blocked by a high building and vehicle, weakening their channel conditions. IRS can play a crucial role in assisting the signal delivery between two vehicles and enhancing their quality of services. IRS can be efficiently installed on the building wall to provide energy-efficient and secure reflection for incident signals toward the desired vehicle. Other vehicles can receive information signals through direct and IRS enabled communications links.

IRS Enabled V2D Communications: The mobility of drones can be efficiently used for communications in densely crowded environments to improve connectivity and reduce the terrestrial network overhead [3]. In particular, large-scale vehicles and other moving objects on roads in big cities can face several issues of performance, connectivity, fading, and transmission latency. It puts an extra burden on the communications network due to the large-scale exchange of data among different vehicles and other objects on the roadside. As shown in Fig. 1, the drone can be operated as a flying BS

[1] While V2X communications encompass numerous use case scenarios, this article mainly focuses on V2I, V2V, V2D, and V2S communications scenarios.

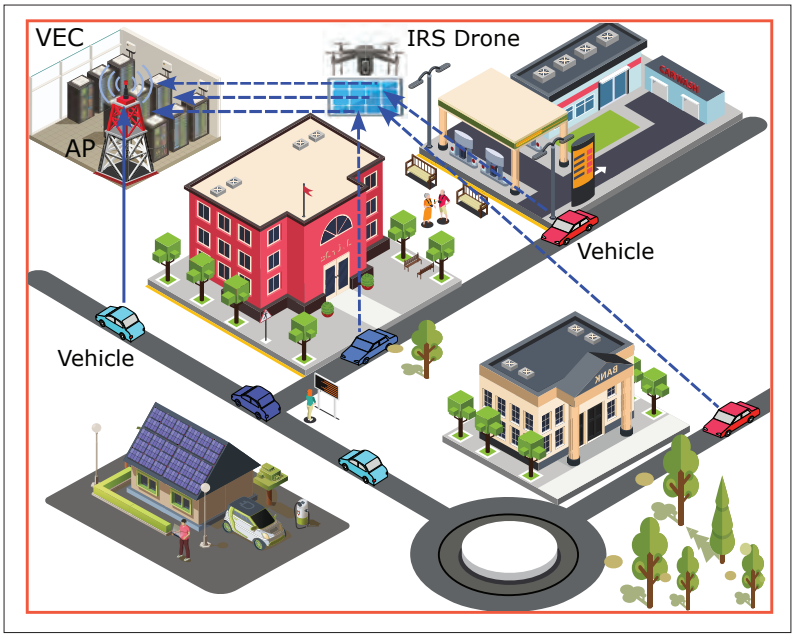


FIGURE 2. System of vehicular edge computing empowered by IRS enabled drone communications.

with the assistance of the IRS to improve the performance of vehicles and other road objects facing NLOS connectivity and poor system performance.

IRS Enabled V2S Communications: Due to the mega low orbit constellation, satellite communications have recently gained significant attention for supporting a wide range of services throughout the globe. It will play a vital role in successfully deploying future autonomous vehicle networks. However, the high mobility of vehicles, obstacles in the urban areas, and shadowing between vehicles and satellites can disrupt the LOS connection and significantly reduce system performance [14]. IRS can be efficiently deployed to deliver signals from satellites to vehicles and improve the link budget. As depicted in Fig. 1, two vehicles on the road are shown to communicate with the satellite, where one vehicle is accessing signal directly from the satellite while the other is facing signal blockage and utilizes IRS to receive signal.

Recent Advances in IRS-Enabled V2X Communications

Although the application of IRS in wireless communications systems has been widely addressed, its application in vehicular environments has been briefly studied in the literature. In the following, we concisely review the most related works and provide their comparison in Table 1.[2]

For example, the work in [4] has introduced IRS and non-orthogonal multiple access (NOMA) in drone enabled wireless communications. The authors simultaneously optimize the phase shift control, dynamic trajectory design, signal decoding order, and power control to minimize the drone's energy consumption. In [5], the authors have considered both communications and computation in a mobile edge computing enabled vehicular network involving IRS. They address the problem of task scheduling which includes allocating processor and link resources. They propose a dynamic task scheduling algorithm to maximize the computation throughput. Resource allocation for IRS enabled V2X communications has also been addressed in [6], where the authors aim to maximize the V2I capacity while guaranteeing the minimum capacity of V2V links. They address a joint power allocation, IRS reflection coefficients, and spectrum allocation problem, proposing an alternating optimization algorithm. To reduce energy consumption and latency, the authors of [7] have proposed a deep reinforcement learning (DRL) strategy for efficient resource allocation in a 3-plane vehicular framework, which includes the vehicle transmission power, IRS reflection phase shift, and BS detection matrix. Moreover, in [8], the authors consider an indirect transmission from roadside units (RSUs) through IRS deployed on buildings to some dark zones. They formulate a joint resource scheduling and RSU passive beamforming aiming at maximizing the minimum average bit rate. In [9], IRS has been integrated into millimetre wave (mmWave) vehicular communications. The authors addressed the issue arising from high mobility and the challenge of obtaining accurate channel state information (CSI). They maximized the average achievable sum-rate by jointly optimizing the transmit power, multi-user detection matrix and the IRS reflection phase shift. A similar problem is studied in [10], where apart from the optimization parameters in [9], they also consider the spectrum reuse of V2V links in the joint formulated mixed-integer non-convex optimization problem. In [11], the authors study a high mobility communications scenario between passengers and BSs through IRS deployed on the vehicle. They address mitigating the Doppler effect through approximations and tuning the IRS reflection over time. In [12], the secrecy outage performance of IRS enabled vehicular communications is analyzed, where they consider two scenarios of V2V and V2I to derive the closed-form expression for the secrecy outage probability. In [13], the authors have maximized the minimum achievable rate of the system by optimizing phase shift design and trajectory by employing binary integer linear programming and soft actor-critic methods.

Resource Optimization for VEC Networks Involving IRS-Enabled Drone Communications

The 6G communications networks would deploy a massive number of sensor nodes to facilitate data sensing from real-time applications. In the context of V2X communications, these sensor nodes can be equipped with vehicles to improve the efficiency of the overall system. These sensor nodes are connected to a central control system known as the access point (AP). However, real-time applications generate a significant volume of data that requires immediate processing. This poses a challenge due to the limited computational resources of the sensor nodes deployed on vehicles. As a result, the performance of V2X communications is compromised. To overcome this challenge, the concept of edge computing has emerged as a promising solution. edge computing enables these devices to offload their data using either a partial or binary offloading scheme. In the partial offloading scheme, tasks are divided into two parts: a portion of the task is computed locally on the sensor nodes, while the other part is offloaded to the edge computing server for extensive computation. This approach alleviates

[2] This article focuses exclusively on vehicular communications, while there are other research papers on the topic of IRS enabled drone communications within conventional networks. Consequently, we have decided not to delve into those works in this study due to space limitations and maintain an appropriate number of references.

Ref.	Use case scenario	IRS position	Transmission	Proposed solution method	Performance gain (objective)
[4]	Drones equipped with multiple antennas communicate with single antenna ground mobile users through IRS	On building	V2I	Decaying deep Q-network	Minimizing energy consumption
[5]	Task offloading of Vehicles to RSU through IRS, where RSU equipped with edge computing	On building	V2I	Dynamic task scheduling algorithm	Maximizing average offloading rate, successfully computing rate, and successfully finish rate
[6]	BS communicates with vehicles through IRS and inter-vehicle communications	On building	V2I & V2V	Alternating optimization	Maximizing sum capacity
[7]	Multiple vehicle clusters, each cluster consists of one head vehicle and multiple member vehicles. Processing requests from head vehicles to BSs through IRS	On building	V2I & V2V	DRL algorithm	Optimizing energy efficiency and latency
[8]	RSU communicates with vehicles lying in dark zone through IRS	On building	V2I	DRL & block coordinate descent algorithms	Maximizing the minimum average bit-rate
[9]	Vehicles communicate with BS through IRS, where BS is equipped with multiple antennas	On building	V2I	Alternating optimization	Maximizing average sum-rate
[10]	Vehicles communicate with BS through IRS and vehicles directly communicate directly with each other	On building	V2V & V2I	BCD algorithm	Maximizing sum capacity
[11]	BS communicates with vehicle through IRS, where IRS is mounted on vehicle	On vehicle	V2I	Heuristic transmission protocol & Passive beamforming	Maximizing achievable rate
[12]	(i) Vehicle communicates with vehicle through IRS in the presence of a passive eavesdropper, and (ii) IRS communicates with the vehicle in the presence of a passive eavesdropper	On building	V2I & V2V	Closed-form expression for secrecy outage probability	Improving secrecy
[13]	BS communicates with high-speed train through direct, and IRS mounted drone links	On drone	V2I	Binary integer linear programming and soft actor-critic methods	Maximizing minimum achievable rate
[Our]	Vehicles communicate with an access point of vehicular edge computing through IRS enable cooperative drone communications	On drone	V2I	Successive convex approximation & standard convex optimization	Reducing the computational time of vehicle task

TABLE 1. Recent advances in IRS enabled V2X communications.

the burden on the sensor nodes and enhances the overall performance of the system. Motivated by these observations, we propose a more practical communications scenario based on VEC network involving IRS enabled cooperative drone communications with aiming to improve the efficiency and effectiveness of the overall system. In the following, we first explain the system model, problem formulation, and the proposed optimization solution. Then we validate our proposed solution by presenting numerical results and their discussion.

System Model, Problem Formulation and Proposed Solution

We consider VEC empowered uplink V2I communications where N low-powered sensor nodes equipped with vehicles are randomly located over a predefined geographical area. All the vehicles are connected to a central control system in VEC called Access Point (AP). We consider that all the nodes are equipped with single antenna scenario. Due to the mobility of vehicles, we consider errors in channel estimation, which is modeled using the minimum mean square error (MMSE) model. According to this model, the modeled channel consists of estimated channel gain and channel errors, where both are assumed to be uncorrelated. In considered VEC system, the AP enables extensive on-demand computation. However, vehicles generate a significant amount of data that necessitates real-time processing, posing a challenge due to their limited computational resources. Consequently, the performance of vehicular networks is compromised. To overcome these limitations, we present a partial offloading scheme that divides tasks into two parts: local computation on the vehicles and offloading to the VEC server for extensive computation. Task segmentation is achieved by analyzing the monotonic relationship between local computation time and offloading time. As the local computation time increases, indicating a larger portion of the task being computed locally, the offloading time decreases. Conversely,

[3] In this work, we assume that the direct link between vehicles and VEC AP is NLOS dominant due to large objects. Thus, we consider an IRS enabled cooperative drone communications to assist transmission between vehicles and VEC AP.

Symbol	Value
Carrier frequency	2 GHz
Vehicle drop model	spatial Poisson process
Path-loss model	128.1+37.6 $\log_{10}(d)$, d in km
VEC AP radius	500 m
Number of IRS reflecting elements	30
Number of vehicles	10
Height of drone	80 m
Bandwidth of the system	20 MHz
Cycles	[2, 10] Kcycles/bits
Data Size	[10, 100] Mbits
Additive white Gaussian noise	−174 dBm
Maximum local computational resources	1 Mcycles/sec
Maximum edge computational resources	25 Gcycles/sec
Velocity of vehicles	60 km/h
Shadowing distribution	Long-normal
Shadowing standard deviation	8 dB
Fast fading	Rayliegh fading

TABLE 2. Simulation parameters and their values [3].

when a significant portion of the task is offloaded to the VEC server, the offloading time increases while the local computation time decreases. Thus, a point is reached where the local computation time becomes equal to the offloading time. This critical point is determined by a mathematical equation called the task offloading percentage, which optimally allocates the task between local computation and offloading.

The task offloading percentage relies on various factors such as local computation resources, edge computation resources, and the achievable data rate over the communication channel. Among these factors, the achievable data rate plays a crucial role, influenced by channel characteristics. In the context of smart cities, high-rise buildings result in higher path loss, leading to increased task offloading time. To address this issue, we consider an IRS enabled cooperative drone communications to provide an alternative wireless communications link, enhancing the channel and received signal gains.[3] Consequently, the achievable data rate of vehicles increases, which significantly reducing the task offloading time. The IRS is mounted over drone and consists of K reconfigurable elements that intelligently manipulate the phase shift of incident signal.

The proposed framework seeks to reduce the computational time of the task in VEC system involving IRS enabled cooperative drone communications. This can be achieved by formulating and solving an optimization problem. In particular, the proposed framework simultaneously optimizes the computational and communications resources among VEC AP and vehicles subject to various practical constraints such as vehicle energy consumption, VEC AP computational resources, drone placement, and efficient phase shift design at IRS, respectively. This framework exploits a partial offloading scheme such that Φ_n percent of the task is computed locally using local resources while the rest is offloaded to the VEC AP for extensive computation. For instance, the amount of computational resources allocated to vehicle n at the local and VEC AP levels can be represented ρ_n^l and ρ_n^e, where $n \in N$. The optimization problem of computational task minimization is formulated as non-convex because of the non-linear objective function. Moreover, coupling decision variables further complicates the optimization problem, and it is very hard to obtain a joint optimal solution. To address the challenges mentioned above and make the optimization more tractable, the original joint optimization problem is divided into multiple subproblems, i.e., drone placement optimization, phase shift control, and computational and communications resources between vehicles and VEC AP, using block coordinate descent (BCD) method. Then, phase shift control and communications resources problems are further transformed using successive convex approximation (SCA) method. The Successive Convex Approximation (SCA) method is valuable for tackling complex optimization problems, especially when dealing with non-convex objectives and constraints. It is an iterative method that can approximate non-convex problems with convex subproblems, making it widely used in wireless communications. After decoupling and transformation, all subproblems are convex, and a standard convex optimization solver, such as CVX, is used to achieve an efficient solution for convex subproblems. Specifically, the proposed method is iteratively updated to find the best possible solution that meets all system constraints. The mathematical method of the related optimization framework can be found in [15].

Numerical Results and Discussion

Here, we provide numerical results of the proposed optimization framework based on Monte Carlo simulations. For optimization, we use the CVX toolbox in MATLAB. The proposed V2I scenario is configured using vehicle drop and mobility according to 3GPP TR 36.885 (see [3] and references therein), where the spatial Poisson process generates vehicles connected to VEC AP, and density is decided through the velocity of the vehicles. Moreover, the channels between IRS enabled drone and infrastructure are established using 3GPP TR 36.777. Unless mentioned otherwise, the simulation parameters are given in Table 2. We compare the following three optimization approaches.

1. *IRS Enabled VHA:* This is the proposed approach involving IRS enabled drone communications, provided earlier, where the tasks are divided into two parts: a portion of the task is computed locally on vehicles, and the remaining part is offloaded to VEC AP.
2. *IRS Enabled VEC:* It refers to a benchmark approach involving IRS enabled drone communications, where all the tasks of vehicles are offloaded to VEC AP.
3. *Without IRS VHA:* It is a benchmark approach without IRS enabled drone communications, in which tasks are partitioned optimally into two portions, one portion of the task is computed locally. At the same time, the other is offloaded to VEC AP without the assistance

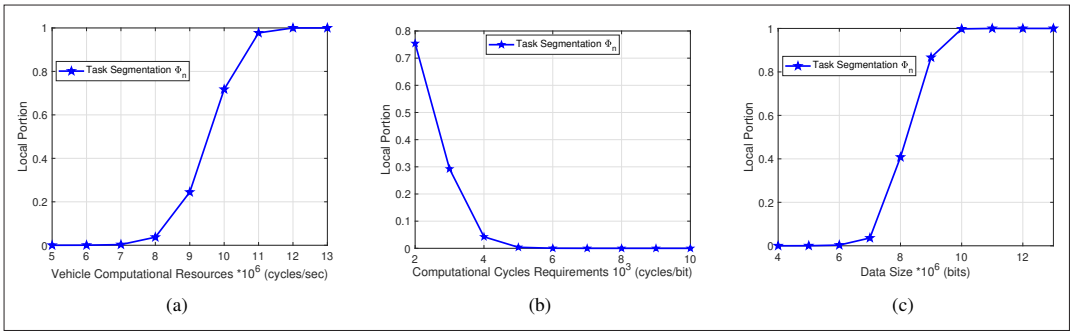


FIGURE 3. Impact of vehicle computational resources, cycles requirement and data size of task offloading.

of IRS enabled drone communications.

In the VEC, task segmentation Φ_n is an important parameter as it directly influences the performance of the system. Φ_n mainly depends on the computational resources of the vehicles, the number of cycle requirements, and the data size of the task, as shown in Fig. 3. In the following, we discuss them one by one.

Impact of Local Computational Resources on Task Segmentation: Computational resources allocated locally play an essential role while determining the task segmentation variables, as shown in Fig. 3a. Results demonstrate that, for the small number of computational resources, the task as a whole is offloaded to VEC AP for extensive computation. This is because local computational resources are not enough to compute it in minimal time. As a result, latency is experienced in a system; hence the quality of service is highly compromised. On the other hand, as the local computational resources increase, the percentage of tasks offloaded decreases, and more tasks are computed locally.

Impact of Computational Cycles Requirements on Task Segmentation: Similarly, computational cycle requirements are also an important parameter while determining the task segmentation parameter. As perceived from Fig. 3a, for the small number of computational cycle requirements, it is efficient to compute the task on the VEC server because of its substantial computational resources. On the other hand, as the computational cycle requirements increase, computational tasks move toward local computation. This trend is because, at the VEC AP, computational cycles are shared among the vehicles, and significant computational cycle requirements demand extensive computation. Whereas shared computation resources at the VEC AP are not enough to meet the high demand; as a result, latency is introduced in a system. To overcome this, local computation is an effective solution to meet the desired requirements, as proved in Fig. 3b.

Impact of Data Size on Task Segmentation: Following that, Fig. 3c demonstrates the impact of data size (bits) on task segmentation. The trend reveals that the task is computed at the VEC AP for the small number of bits. As the data size increases, the computational task shifts toward a local computational scheme. This trend is because offloading a small number of bits consumes less offloading energy than a task's computational energy locally. In contrast, as the data size increases, the offloading energy consumption is more than the local computational energy. So it is efficient to compute the task locally.

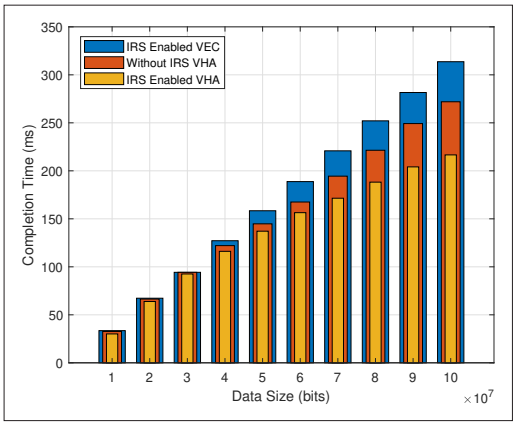


FIGURE 4. Impact of IRS on task's completion time.

Next, Fig. 4 represents the impact of task segmentation and communications links on the system's performance. The results reveal the proposed *IRS Enabled VHA* approach outperforms the others by considering computational task time as a performance metric. For the small number of data sizes, the performance of *IRS Enabled VEC* closely follows the other approaches. However, as the data size requirements increase, the proposed *IRS Enabled VHA* approach and *Without IRS VHA* start performing better. This trend is due to the fact that offloading a large number of bits constitutes more time and energy than the *IRS Enabled VEC* approach. Whereas in the proposed *IRS Enabled VHA* and *Without IRS VHA* approaches, the task is divided into two portions optimally. One portion of the task is computed locally, whereas the other is offloaded to VEC AP. Likewise, offloading a portion of the task also constitutes less time and energy than complete offloading, compared to *IRS Enabled VEC* approach.

In addition, the results also demonstrate the IRS's impact on the system's performance. Comparative analysis of *IRS Enabled VEC* and *Without IRS VHA* further reveal the effectiveness of IRS in the VEC network. *Without IRS VHA* is a traditional approach in which users offload their portion of the task to VEC AP for extensive computation. At the same time, the presence of an obstacle in the paths results in a weakness in the received signal strength. Therefore, reducing the achievable data rate requires more time to offload the task to VEC AP. Hence, latency is introduced into the system. Whereas in *IRS Enabled VEC* approach, the drone is equipped with IRS to assist the communications between vehicles and VEC AP, resulting strong signal that helps minimize the offloading time. Overall computational time decreases as a result.

This trend is because, at the VEC AP, computational cycles are shared among the vehicles, and significant computational cycle requirements demand extensive computation.

Open Issues and Future Research Directions

In this section, we discuss and highlight all open issues and future research directions.

Machine Learning Techniques

In future 6G systems, numerous machine learning (ML) techniques are envisioned to perform intrinsic and complicated tasks related to resource allocation and signal processing. In this regard, ML techniques offer a competitive edge over traditional methods or algorithms for seamlessly performing computationally intensive tasks. %Moreover, only a handful of works in open technical literature have considered enabling V2X communications via ML techniques despite their wide availability and unlimited capabilities. Furthermore, research on facilitating IRS enabled V2X communications via ML techniques is a viable futuristic research direction. However, it has not come under the limelight and has not received much attention from the research community.

In this context, ML techniques could solve various optimization, classification, prediction, and decision-making tasks which could be extremely useful for facilitating IRS enabled V2X communications.

Drone IRS Enabled V2X Communications

The upcoming 6G systems will support non-terrestrial networks based on drones serving as aerial BSs or relays. According to 3GPP Release 15, a drone flying at an altitude of 80 meters or above has a 100% probability of achieving LOS communications capable of mitigating shadowing and signal blockage. IRS can be efficiently mounted on drones to satisfy the QoS requirements of vehicles, such as ultra-reliable and low-latency communications. Therefore, its deployment in V2X communications is a promising research direction to investigate.

Power Consumption and Energy Efficiency

IRS requires powering a large number of elements to manipulate the electromagnetic environment. Optimizing performance while managing power consumption is essential, especially in vehicle environments with limited resources. Research is still being done on developing power allocation algorithms and energy-efficient IRS systems.

Optimal Beamforming

Beamforming optimization in IRS enabled terrestrial and non-terrestrial networks is challenging. In drone-based non-terrestrial networks, wind gusts can cause random jittering, which can miscalculate the IRS and vehicle angle of departure. Most IRS studies assume continuous phase shifts, which is difficult owing to hardware constraints. Signal misalignment and IRS beamforming may result from this assumption. Beamforming optimization with discrete phase shifts at IRS in V2X communications is essential and a promising area of study to address this problem.

Scalability and Complexity

The deployment of IRS should be scalable to handle the growing number of users because V2X systems frequently involve a large number of vehicles. Large-scale IRS implementation, however, adds complexity to signal processing, synchronization, and control. Creating scalable and manageable designs to handle the complexity is an immense challenge that should be addressed.

mmWave and THz Communications

High-traffic vehicle networks require mmWave and THz frequency to meet user demand. Data rates of 10 Gb/s across 850 m have been achieved at 120 GHz. Transceivers beyond 300 GHz need power, sensitivity, and low noise to overcome the issue of a high path loss. mmWave and THz V2X communications are restricted by penetration losses, significant Doppler dispersion, and blockage. IRS can overcome such challenges and contend for V2X communications in mmWave and THz frequencies. Mutual coupling and electromagnetic interference make IRS development for mmWave and THz bands a viable research topic.

Interference Mitigation

Interference is a big challenge in heterogeneous V2X scenarios where different vehicles simultaneously share the same spectrum. IRS can be applied to manage the reflected signals, reducing interference actively. While multiple vehicles and IRS elements coexist in V2X environments, minimizing interference is a challenging problem. Effective interference management algorithms and tactics are needed to maximize system performance and guarantee reliable V2X communications.

Standardization and Deployment

Although IRS technology has a lot of potentials, some practical hurdles must be overcome before it can be practically used in V2X communications systems. These difficulties include standardization of IRS enabled V2X networks, cost-effectiveness, regulatory issues, and integration with current infrastructure.

Conclusion

The IRS has been regarded as an emerging technology in 6G, with the goal of controlling signal propagation and creating a smart radio environment. The IRS is designed to provide LOS-like propagation and energy-efficient communications in 6G V2X networks by leveraging intelligent reflection capabilities. The IRS can help V2X communications in 6G networks by improving multipath propagation and increasing transmission coverage in high spectrum situations. This article discussed the potential and opportunities of the IRS in 6G empowered V2X communications. In particular, we described different use case scenarios in IRS enabled V2X communications and discussed recent advances. Then, we provided a case study on resource optimization of VEC network involving IRS enabled cooperative drone communications. The numerical results showed the benefits of IRS in terms of task computational time. Finally, we also highlighted current issues and some potential research directions in IRS enabled V2X communications.

References

[1] S. Gyawali *et al.*, "Challenges and Solutions for Cellular Based V2X Communications," *IEEE Commun. Surveys & Tutorials*, vol. 23, no. 1, 2020, pp. 222–55.

[2] W. U. Khan *et al.*, "NOMA-Enabled Backscatter Communications for Green Transportation in Automotive-Industry 5.0," *IEEE Trans. Industrial Informatics*, vol. 18, no. 11, Nov. 2022, pp. 7862–74.

[3] A. Ihsan *et al.*, "Energy-Efficient NOMA Multicasting System for Beyond 5G Cellular v2x Communications with Imperfect

csi," *IEEE Trans. Intell. Transportation Systems*, vol. 23, no. 8, 2021, pp. 10,721–35.
[4] X. Liu, Y. Liu, and Y. Chen, "Machine Learning Empowered Trajectory and Passive Beamforming Design in UAV-RIS Wireless Networks," *IEEE JSAC*, vol. 39, no. 7, 2021, pp. 2042–55.
[5] Y. Zhu, B. Mao, and N. Kato, "A Dynamic Task Scheduling Strategy for Multi-Access Edge Computing in IRS-Aided Vehicular Networks," *IEEE Trans. Emerging Topics in Computing*, 2022, pp. 1–1.
[6] Y. Chen *et al.*, "Resource Allocation for Intelligent Reflecting Surface Aided Vehicular Communications," *IEEE Trans. Vehic. Tech.*, vol. 69, no. 10, 2020, pp. 12,321–26.
[7] Q. Pan *et al.*, "Artificial Intelligence-Based Energy Efficient Communication System for Intelligent Reflecting Surface-Driven VANETs," *IEEE Trans. Intelligent Transportation Systems*, 2022, pp. 1–13.
[8] A. Al-Hilo *et al.*, "Reconfigurable Intelligent Surface Enabled Vehicular Communication: Joint User Scheduling and Passive Beamforming," *IEEE Trans. Vehic. Tech.*, vol. 71, no. 3, 2022, pp. 2333–45.
[9] Y. Chen, Y. Wang, and L. Jiao, "Robust Transmission for Reconfigurable Intelligent Surface Aided Millimeter Wave Vehicular Communications with Statistical CSI," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, 2022, pp. 928–44.
[10] Y. Chen *et al.*, "QoS-Driven Spectrum Sharing for Reconfigurable Intelligent Surfaces (RISs) Aided Vehicular Networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, 2021, pp. 5969–85.
[11] Z. Huang, B. Zheng, and R. Zhang, "Transforming fading Channel from Fast to Slow: IRS-Assisted High-Mobility Communication," *ICC 2021 — IEEE Int'l. Conf. Commun.*, 2021, pp. 1–6.
[12] Y. Ai *et al.*, "Secure Vehicular Communications Through Reconfigurable Intelligent Surfaces," *IEEE Trans. Vehic. Tech.*, vol. 70, no. 7, pp. 7272–7276, 2021.
[13] Y. M. Park *et al.*, "Trajectory Optimization and Phase-Shift Design in IRS-Assisted UAV Network for Smart Railway," *IEEE Trans. Vehic. Tech.*, vol. 71, no. 10, Oct. 2022, pp. 11,317–21.
[14] W. U. Khan *et al.*, "Opportunities for Physical Layer Security in UAV Communication Enhanced with Intelligent Reflective Surfaces," *IEEE Wireless Commun.*, vol. 29, no. 6, Dec. 2022, pp. 22–28.
[15] A. Mahmood *et al.*, "Optimal Resource Allocation and Task Segmentation in IoT Enabled Mobile Edge Cloud," *IEEE Trans. Vehic. Tech.*, vol. 70, no. 12, Dec. 2021, pp. 13,294–03.

Biographies

Wali Ullah Khan [M] (waliullah.khan@uni.lu) received a Ph.D. degree in information and communication engineering from Shandong University, Qingdao, China, in 2020. He is currently working with the SIGCOM Research Group, SnT, University of Luxembourg.

Asad Mahmood [S] (asad.mahmood@uni.lu) received his Master degrees in Electrical Engineering from COMSATS University Islamabad, Wah Campus, Pakistan. He is currently pursuing the Ph.D. degree with the Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg.

Arash Bozorgchenani [M] (a.bozorgchenani@leeds.ac.uk) received a Ph.D. degree in Telecommunications and IT from the University of Bologna in Italy, in 2020. He is currently an Assistant Professor with the School of Computing, University of Leeds, UK.

Muhammad Ali Jamshed [SM] (muhammadali.jamshed@glasgow.ac.uk) received a Ph.D. degree from the University of Surrey, Guildford, U.K, in 2021. He is currently working with the James Watt School of Engineering, University of Glasgow, UK.

Ali Ranjha (ali-nawaz.ranjha.1@ens.etsmtl.ca) received a Ph.D. degree in Engineering from Ecole de Technologie Superieure (ETS), Universite du Quebec, Montreal, Canada in January 2022, where he is currently pursuing his postdoctoral research.

Eva Lagunas [SM] (eva.lagunas@uni.lu) received a Ph.D. degree in telecommunications engineering from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, in 2014. She currently holds a research scientist position in the SIGCOM Research Group, SnT, University of Luxembourg.

Haris Pervaiz [SM] (h.b.pervaiz@lancaster.ac.uk) received a Ph.D. degree from Lancaster University, U.K., in 2016. He is currently an associate professor at the School of Computer Science and Electronic Engineering, University of Essex, UK.

Symeon Chatzinotas [F] (symeon.chatzinotas@uni.lu) received Ph.D. degrees in electronic engineering from the University of Surrey, Guildford, United Kingdom, in 2009. He is currently a full professor or Chief Scientist I and the co-head of the SIGCOM Research Group, SnT, University of Luxembourg.

Björn Ottersten [F] (bjorn.ottersten@uni.lu) received his Ph.D. degree in electrical engineering from Stanford University, California, in 1990. He is currently the director for SnT, University of Luxembourg.

Petar Popovski [F] (petarp@es.aau.dk) received his Ph.D. degree from Aalborg University in 2005. He is currently a full professor with Aalborg University, Denmark.

ISAC-Assisted Wireless Rechargeable Sensor Networks with Multiple Mobile Charging Vehicles

Muhammad Umar Farooq Qaisar, Weijie Yuan, Paolo Bellavista, Guangjie Han, and Adeel Ahmed

Abstract

As IoT-based wireless sensor networks (WSNs) become more prevalent, the issue of energy shortages becomes more pressing. One potential solution is the use of wireless power transfer (WPT) technology, which is the key to building a new shape of wireless rechargeable sensor networks (WRSNs). However, efficient charging and scheduling are critical for WRSNs to function properly. Motivated by the fact that probabilistic techniques can help enhance the effectiveness of charging scheduling for WRSNs, this article addresses the aforementioned issue and proposes a novel ISAC-assisted WRSN protocol. In particular, our proposed protocol considers several factors to balance the charging load on each mobile charging vehicle (MCV), uses an efficient charging factor strategy to partially charge network devices, and employs the ISAC concept to reduce the traveling cost of each MCV and prevent charging conflicts. Simulation results demonstrate that this protocol outperforms other classic, cutting-edge protocols in multiple areas.

Introduction

The Internet of Things (IoT) has revolutionized the way we interact with technology, from smart homes to wearable devices [1]. Central to this transformation are wireless sensor networks (WSNs), which enable the connection and information transmission of devices and systems. However, WSNs face a significant challenge in the form of an energy shortage, which can impact their capability to function effectively [2]. This is where the concept of wireless rechargeable sensor networks (WRSNs) comes in, utilizing wireless power transfer (WPT) technology to ensure energy sustainability. WPT is employed to wirelessly recharge the energy-starved sensor devices. Three popular WPT technologies are inductive coupling, electromagnetic (EM) radiation, and magnetic resonant coupling. In contrast to the initial two techniques, the magnetic resonant coupling exhibits superior energy transfer efficiency under omni-direction, eliminates the need for line-of-sight (LOS), and is unaffected by external factors. WRSNs have practical applications in various fields of IoT, including smart cities, smart healthcare, smart farming, smart traffic, and smart homes. Typically, a WRSN is comprised of a base station, which serves as a depot for mobile charging vehicles (MCVs), one or more MCVs, and sensor devices equipped with rechargeable batteries that receive wireless signals to recharge from MCVs as depicted in Fig. 1. This unique charging approach enables WRSNs to operate efficiently and continuously, without interruption [3].

To recharge the sensor devices in WRSNs, the MCVs use either periodic charging or on-demand charging strategies. While periodic charging follows a predetermined schedule, it is not always ideal due to the dynamic energy depletion rate of the sensor devices. In contrast, on-demand charging is more flexible and is capable of making real-time decisions based on the energy requirements of the sensor devices. Additionally, charging strategies can be either full or partial charging models. Full charging results in significant charging delays, whereas partial charging allows for more sensor devices to be recharged. Nevertheless, existing strategy designs fail to take into account the traveling time and potential conflicts between multiple MCVs, which are also of great importance in the charging process [4].

We intend to address the aforementioned problem by utilizing the innovative Integrated Sensing and Communication (ISAC) technique [5]. By integrating the functionalities of sensing and communications, this technique allows for efficient utilization of wireless resources, wide-area environmental sensing, and mutual benefits. Moreover, by exploiting the benefits of wireless signals [6], ISAC can improve the charging efficiency of MCVs and reduce their travel time. In this work, the charging sensor devices can be arranged in a queue, with multiple MCVs having different priorities. When an MCV comes within range of a prioritized charging sensor device, the sensor device transmits an ISAC signal to the vehicle. After receiving an echo of the signal via wireless transmission, the sensor device analyzes it and communicates with the base station to update the charging priorities of other MCVs in the queue. This approach improves the efficiency

Muhammad Umar Farooq Qaisar (corresponding author) is with Northwestern Polytechnical University and Southern University of Science and Technology, China; Weijie Yuan (corresponding author) is with Southern University of Science and Technology, China; Paolo Bellavista is with the University of Bologna, Italy; Guangjie Han is with Hohai University, China; Adeel Ahmed is with the University of Science and Technology of China, China.

Digital Object Identifier: 10.1109/IOTM.001.2300153

2576-3180/24/$25.00 © 2024 IEEE

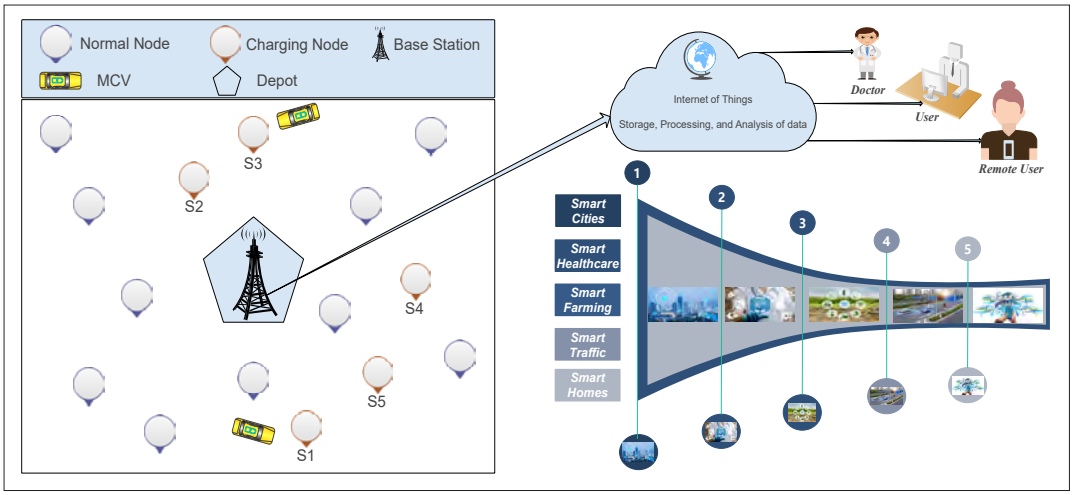


FIGURE 1. Wireless rechargeable sensor networks for IoT applications.

of charging by minimizing travel time and avoiding conflicts that might arise when multiple MCVs attempt to charge the same sensor device.

The studies mentioned above have identified several compelling reasons for addressing the challenges associated with deploying multiple MCVs and developing an effective on-demand charging strategy for sensor devices, as well as integrating the ISAC concept with WRSNs to optimize network stability. Thus, this article introduces the ISAC-assisted WRSNs protocol, which includes three key strategies. The first strategy ensures a balanced charging load across each MCV priority queue by taking into account four attributes: residual energy of the charging device, the distance between the MCV and the charging device, the degree of the charging device, and the charging device betweenness centrality. The second strategy determines the charging factor for each MCV queue, enabling the partial charging of all sensor devices to further enhance charging efficiency and network lifetime. Finally, the third strategy integrates the ISAC concept with WRSNs to leverage wireless resources, minimize travel time, and avoid conflicts that may occur when multiple MCVs seek to charge the same sensor device. The main goal of this study is to propose a novel charging strategy for establishing a balanced load distribution across several MCVs using a highly effective on-demand charging technique. This strategy will result in noticeable improvements in the overall effectiveness of charging. In addition, the study presents a cutting-edge sensing and communication technique that significantly reduces the amount of time MCVs spend within the network. The findings of our developed protocol show that it outperforms more recent state-of-the-art protocols in terms of performance and provides strong evidence of its ability to improve MCV charging efficiency while decreasing travel time.

Related Work

The significance of energy replenishment in WRSNs grows as we progress deeper into the IoT domain. This section briefly examines the studies on WRSN energy replenishment that are pertinent to our work and sheds light on the most recent developments and cutting-edge approaches in this field.

The authors of [7] state in a paper that they have developed a charging method that clusters each device's energy requirements in order to provide an equal distribution of the charging load across MCVs. This technique significantly increases the number of recharged devices while decreasing charging time. To reduce charging delays, a charging scheduling mechanism is presented in [9]. The authors set up a closed charging tour for each MCV in an effort to stop sensor devices from being charged by multiple MCVs at once. Their method, meanwhile, led to an uneven distribution of charging loads among the MCVs. A distributed mobile charging methodology is proposed in [10] and is intended to schedule multiple MCVs in congested WRSNs. The authors utilized game theory techniques to tackle the issue of multi-charging and employed an on-demand partial charging strategy that resulted in a repetitive game played by the MCVs. The authors claimed that their method resulted in better charging coverage and a decrease in charging time. In [11], the authors tackled the challenge of coordinating multiple MCVs to schedule and optimize charging with the aim of reducing the overall energy consumption of the MCVs. Their approach involved modifying the mobility speed and charging time to minimize travel time and improve efficiency. The study presented by the authors in [12] is an uneven cluster-based mobile charging method that divides sensor devices into groups and optimizes the charging schedule for each MCV based on residual energy and distance to sensor devices. However, their method leads to reduced charging efficiency. A study in [13] proposed a charging scheduling approach that uses fuzzy logic to manage multiple MCVs. The authors aimed to equally share the charging load of sensor devices among the MCVs by dividing the network. They also established dynamic charging thresholds for each sensor device based on its rate of energy consumption. Additionally, the authors used fuzzy logic and multi-metric inputs to determine the next sensor device to be charged for each MCV. Their approach falls short in efficiently selecting the next sensor device to be charged due to the inefficiency of the multi-metric strategy employed. A recent work in [14] developed a novel approach to minimize charging

The findings of our developed protocol show that it outperforms more recent state-of-the-art protocols in terms of performance and provides strong evidence of its ability to improve MCV charging efficiency while decreasing travel time.

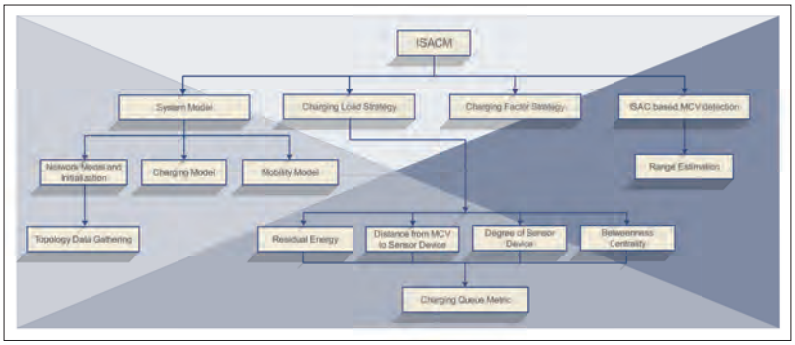


FIGURE 2. The core architecture of the proposed novel scheme.

delays in WRSNs through charging sensor devices with multiple MCVs. This method, in contrast to earlier works, used predetermined MCV travel trajectories with varying speeds.

Previous research on charging scheduling strategies in WRSNs has demonstrated their feasibility and potential usefulness. These approaches, however, lacked effectiveness in balancing charging loads among MCVs and implementing a viable charging factor strategy for partially charging network devices. Furthermore, they were inadequate in addressing the problem of minimizing travel time for multiple MCVs and resolving charging conflicts in the network. To address these issues, this article develops a novel scheme that employs an effective multi-metric and charging factor strategy while also introducing the concept of ISAC to WRSNs. This concept optimizes the sensing and communication tasks between charging devices and MCVs, resulting in reduced travel time and charging conflicts. This approach goes beyond the current state-of-the-art WRSN charging scheduling strategies.

System Model

Network Model and Initialization

With the assumption of a wireless rechargeable sensor network, we have a set of randomly deployed sensor devices and a set of MCVs in a two-dimensional region. When the energy level of the sensor devices drops below a specific threshold, they send charging requests to the base station, which is positioned at the central location within the region and serves as a depot for the MCVs. This research adapts the initial position of the MCVs, described in [9], to determine the coordinates of the MCVs' initial position in k different regions. At a time, the MCV recharges only one sensor device. Obviously, the energy capacity of an MCV is significantly higher than that of a sensor device.

Charging Model

According to the WPT technique, the sensor device can only receive a fraction of the power transmitted by the MCV. The WPT efficiency between an MCV and a sensor device is non-linearly correlated with the distance between the two, indicating that the charging efficiency decreases as the distance increases. The charging process is considered successful only when the MCV arrives at the precise location of the sensor device. Therefore, we aim to achieve the desired efficiency by employing a partial-charging strategy, and the requesting sensor devices may not be fully recharged. It is important to note that the effectiveness of WPT is consistent across all sensor devices. However, MCV has limited energy capacity and must return to the sink for self-recharging when its residual energy falls below a predetermined threshold. The base station serves as a depot for the MCVs.

Mobility Model

We adopt a recent and widely accepted extension of the Random Waypoint (RW) mobility model, which has been used in a recently published state-of-the-art article [14]. The charging time of the sensor devices is described by the pause value at each stopover location and the positioning plan based on the direction of the subsequent sensor devices in the queue awaiting charging. We assume that the MCV travels along a straight, obstacle-free path at a constant speed, and its location is continually updated until it reaches the location of the next sensor device in line to be charged. When the distance between the MCV and the charging sensor device is less than or equal to the distance threshold for efficient energy transmission, the charging process takes place.

The Proposed Protocol

In this article, we propose a technique for balancing the charging load on each MCV and partially charging all sensor devices in a queue in order to improve charging efficiency. We also incorporate the ISAC technique to make use of wireless resources to reduce the network travel time of the MCVs. This strategy can be applied to potential IoT applications where network coverage and energy efficiency are important considerations. Figure 2 displays the basic design of the proposed *ISACM*.

Charging Load Strategy

In this section, we present a strategy for balancing the charging load for each MCV queue in the wireless rechargeable sensor network. To accomplish this, we consider four important attributes: residual energy of the charging sensor device, distance from MCV to charging sensor device, degree of a charging sensor device, and charging sensor device betweenness centrality. Each attribute is associated with a probability distribution function, and the charging queue metric is determined by taking their respective values into account. The charging sensor devices with the highest value of this metric are prioritized in the charging queue. Because the MCVs are distributed in k different locations throughout the network, each MCV will have a distinct sequence of prioritized charging sensor devices.

Residual Energy of Charging Sensor Device: In the context of IoT networks, optimal scheduling of charging sensor devices is essential with residual energy being an important factor to consider. This section explains how to efficiently use residual energy as a scheduling variable to prioritize the charging process. The attribute reflects a charging sensor device's residual energy below the residual energy threshold and prioritizes sensor devices with lower residual energy than other charging devices.

Distance from MCV to Charging Sensor Device: In IoT networks, it is essential to effectively schedule charging sensor devices, and the distance between MCV and charging sensor devices must be taken into account. Sensor devices will be prioritized according to how far they are from the MCV with the help of this attribute. The devic-

es with the shortest distance will be given the highest level of priority.

Degree of Charging Sensor Device: The degree of a charging sensor device in an IoT network, which can be estimated by the number of connections it has with adjacent devices, is essential to the data flow rate of that device.

A higher number of neighbors results in a greater probability of higher data flow to the charging sensor device. Moreover, if a device has low energy levels below the charging threshold, a relatively high data flow rate from numerous neighbors could quickly deplete its residual energy, thereby making efficient scheduling highly demanded. Therefore, this attribute aims to prioritize sensor devices with the maximum number of neighbors among all the charging sensor devices.

Charging Sensor Device Betweenness Centrality: The concept of betweenness in an IoT network refers to the number of times a sensor device acts as a bridge between two other devices along the shortest path. As such, the more frequently a charging sensor device acts as a betweenness, the greater the chance of information flow, which can simultaneously deplete its energy and make it critical. Therefore, the purpose of this attribute is to prioritize charging sensor devices that frequently serve as bridges.

The attributes are normalized within the range of 0 and 1. Based on this normalization, we determine the probability distribution function of each attribute through curve fitting. Additionally, we introduce a weighted factor for each attribute to augment the influence of the probability distribution functions and prioritize charging sensor devices by increasing their respective values.

Charging Queue Metric: The objective of the charging queue metric is to effectively prioritize the charging queue of each MCV by utilizing the four aforementioned attributes. As outlined earlier, the initial location of each MCV is partitioned into k regions. Therefore, depending on the distance attribute between the MCV and the sensor device, the priority of charging sensor devices in each MCV queue will vary. In this work, each MCV queue's priority is determined by the base station using the charging queue metric. The average value of all four attributes is used to calculate the charging queue metric, which is given the highest priority value.

Charging Factor Strategy

The probabilistic partial charging model presented in this section focuses on the charging factor strategy to increase charging efficiency by taking into account the criticality of charging sensor devices based on their residual energy attributes. The probabilistic weighted factor plays an vital role in determining the order in which charging tasks should be completed, and the presence of the charging control factor ensures fair and effective partial charging, improving charging efficiency, coverage, and longevity of the network. The charging factor strategy is depicted in detail in Fig. 3.

ISAC-Based MCV Detection

In the context of WRSN, the travel costs incurred by MCVs are directly related to ISAC. To increase the overall efficiency of WRSNs, ISAC implies the mutual improvement of sensing and communication capabilities. When it comes to MCV travel costs, ISAC plays an essential role in maintaining the balance between the relationship between travel costs, travel time, and rewards. To increase the energy usage efficacy of wirelessly charging sensors, ISAC attempts to decrease the cost of travel. Thus, the relationship between ISAC and MCV travel costs in WRSNs is based on optimizing energy usage for recharging and data collection, which are critical to the network's optimal operation.

FIGURE 3. The figure illustrates the steps involved in the charging factor strategy: 1) It starts with the charging sensors in the MCV queue. 2) The residual energy attribute of the charging sensor device is taken into consideration in the queue. 3) The residual energy values are then used to calculate the criticality values, representing the minimum and maximum criticality of the sensor devices. 4) These criticality values are then utilized in the calculation of the probabilistic weighted factor. 5) The charging control factor is obtained for each sensor device in the queue to regulate the charging process efficiently, adjusted to 10% of the residual energy priority for each sensor device. 6) Finally, the charging factor strategy is determined based on the above factors to enhance charging efficiency and ensure fair and successful partial charging.

In the ISAC approach, the charging sensor device in the prioritized queue transmits an ISAC signal to the closest MCV within its sensing range. With the use of the ISAC signal, the device can determine distance by analyzing the received echo signal while accounting for noise and interference. Figure 4 illustrates how the device then engages in communication with the base station to avoid other MCVs from approaching and overcharging it.

This cuts down on the time it takes for MCVs to charge and their travel time. By evaluating the information collected using this method, it enables the base station to modify the priority of other MCVs' charging queues.

The objective is to extract information from a cluttered received signal because both noise and interference affect the received echo signal. In order to do this, the sensor device applies matched filtering to the signal it receives. The matched filter is specifically designed to increase the correlation between the received signal and the signal sent by the MCV.

To achieve the maximum cross-correlation function time delay, we begin by rewriting the cross-correlation function as a function of time delay and then compute the derivative of the cross-correlation function with respect to time delay. The optimal time delay that maximizes the cross-correlation function is the one that maximizes the convolution of the received echo signal and conjugate of the signal transmitted by the MCV.

The strategy for the charging factor involves using the charging control factor, which aims to optimize the accessibility of each sensor device in the queue.

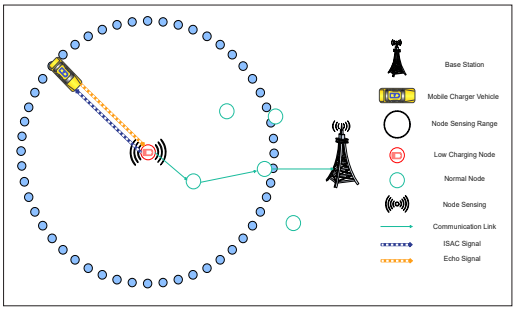


FIGURE 4. MCV detection based on ISAC approach.

Parameter	Value
Number of sensor devices	Varies from 100 to 500
Communication range	50*m*
Sensing range	25*m*
Sensor device battery capacity	0.5*J*
Threshold for charging requests	30% residual energy
MCV battery capacity	10*kJ*
The charging rate	0.05*J*/*s*
MCV travel speed	5*m*/*s*
MCV travel cost	5*J*/*m*

TABLE 1. Simulation parameters.

The charging sensor device estimates the distance to the MCV by maximizing the cross-correlation function with a time delay and using the speed of light to calculate the distance. If the distance is within the sensing region, the sensor device detects the MCV and communicates with the base station according to the probabilistic sensing model [15]. Following that, the base station removes that sensor device and updates the priority queues of other MCVs.

Performance Evaluation and Discussions

In order to evaluate the performance of the proposed protocol, a carefully designed simulation setup was employed to conduct comprehensive simulations. The WRSN network is set up in a square monitoring area, with sensor devices distributed at random. The base station, which is positioned in the middle of the area, acts as a depot where MCVs may replenish their batteries. The scheduling of charging requests for each MCV and network management are also its responsibilities. We performed 20 random simulations, averaged the results, and verified the accuracy of our results. Table 1 presents a summary of the important simulation parameters and their corresponding values.

The performance of the proposed protocol is measured based on several key parameters. These parameters include:

- **Energy Usage Efficiency:** This parameter is defined as the ratio of the total energy transferred to the sensor devices to the total energy transmitted from the base station to the MCVs.
- **Charging Delay:** This parameter is defined as the time it takes for the MCVs to fulfill the energy requirements of the sensor devices.
- **Travel Distance:** This parameter is defined as the total distance covered by the MCV during a single charging tour.

The proposed protocol (*ISACM*) is compared with two recent and state-of-the-art protocols, namely DMCP [9] and FLCSD [12], to assess its effectiveness.

The energy usage efficiency results for the proposed protocol, FLCSD, and DMCP are depicted in Fig. 5a. The proposed protocol outperforms the state-of-the-art protocols due to several factors. Firstly, it employs a charging load strategy that prioritizes charging sensor devices in each MCV queue fairly. The strategy is based on a probability distribution, ensuring a balanced allocation of charging load. Secondly, an effective charging factor strategy is employed to partially charge all the sensor devices in the network. This strategy is based on the residual energy of the charging sensor device attribute. Finally, it maximizes the energy transferred to each requested sensor device in the network by employing the ISAC concept, which reduces the travel cost of each MCV. However, both FLCSD and DMCP protocols lacked efficient charging loads for each MCV in the network and prioritized charging queues, resulting in suboptimal energy usage efficiency.

The charging delay results are depicted in Fig. 5b, showing a gradual increase in delay with the number of sensor devices for all protocols. The proposed protocol achieves better performance compared to existing protocols by utilizing a probabilistic partial charging approach that covers more sensor devices in each MCV queue. In this method, a charging factor is assigned to each sensor device in the MCV queue based on its level of criticality, which assists in effectively partially charging the sensor devices. The strategy for the charging factor involves using the charging control factor, which aims to optimize the accessibility of each sensor device in the queue. Furthermore, the protocol adopts the ISAC concept to minimize travel costs and enhance the likelihood of charging the intended sensor device in an efficient manner. FLCSD protocol did not incorporate the partial charging approach. In contrast, DMCP protocol considered only the relative criticality and did not utilize a charging factor strategy that involves a probabilistic approach and charging control factor to decrease charging delay.

Figure 5c illustrates the variation in travel distance with an increase in the number of sensor devices for all protocols. The proposed protocol exhibits superior performance compared to existing protocols as it prioritizes sensor devices in the charging queue that are closer to the MCV using the distance from the MCV to the charging sensor device attribute. Additionally, it adopts the ISAC approach as described in the charging delay result. In contrast, FLCSD and DMCP protocols did not consider minimizing the travel distance of the MCV in the network.

Conclusion

The presented work proposes an efficient charging solution for IoT applications through the use of ISAC-Assisted WRSNs with Multiple MCVs. The work centers around three key areas: load balancing charging scheduling for each MCV, an efficient charging factor strategy for partial charging of network devices, and an integrated sensing and communication approach to reduce the travel costs of MCVs. To accomplish this objective, the work initially evaluates four attributes with their corresponding probability distribution functions to balance the charging load and prioritize crucial sensor devices for charging. These attributes are the residual energy of the charging

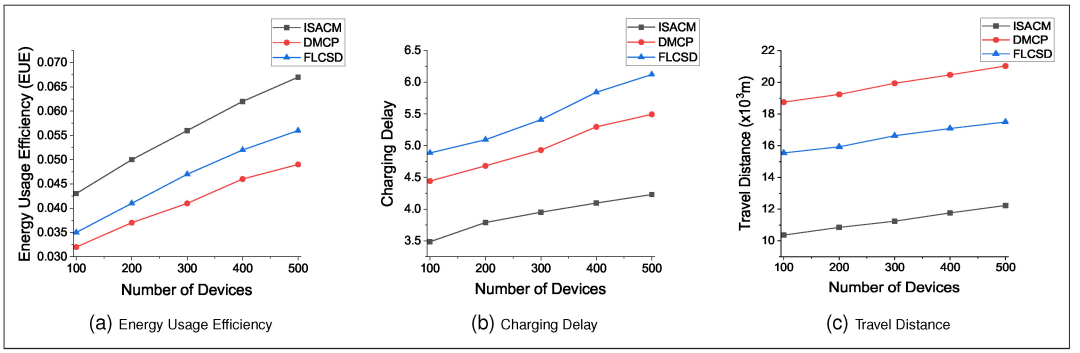


FIGURE 5. Performance over number of devices.

sensor device, the distance from MCV to the charging sensor device, the degree of charging sensor device, and the charging sensor device betweenness centrality. Next, it introduces an efficient strategy for charging that is based on the probability distribution function of residual energy of the charging sensor devices. Additionally, a charging control factor is included to minimize charging delays while increasing charging coverage. Lastly, the article utilizes the ISAC concept to identify the nearest MCV that arrives within the sensing range of the charging sensor device. This approach decreases the travel distance of other MCVs in the network and avoids charging conflicts between them. According to the simulation results, the proposed protocol exhibits superior performance compared to existing protocols.

Future Directions and Challenges

This section discusses potential future research directions as well as the current challenges in implementing ISAC-assisted WRSNs with multiple MCVs. It discusses the need for advancements in this area to improve the performance of WRSNs, as well as the open problems and challenges that still need to be addressed. This section aims to open the door for the development of effective charging strategies for IoT applications within WRSNs by addressing these future directions and challenges.

Future Research Directions

In the context of ISAC-assisted WRSNs with multiple MCVs, several promising research directions can be explored to further enhance the efficiency and effectiveness of the charging solution for IoT applications. First, the integration of machine learning algorithms holds great potential for optimizing the performance of such networks. By leveraging machine learning techniques, researchers can develop intelligent algorithms that adaptively control the charging operations, optimize resource allocation, and predict the charging demands based on historical data, network conditions, and user behavior patterns.

Another important research direction is the investigation of advanced energy harvesting techniques. Exploring novel methods such as solar, kinetic, or RF energy harvesting can significantly contribute to the self-sustainability of WRSNs by harnessing ambient energy sources. This would not only extend the lifetime of the sensor devices but also reduce the dependence on external power sources and increase the overall energy efficiency of the network.

Open Problems and Challenges

While ISAC-assisted WRSNs with multiple MCVs offer promising charging solutions for IoT applications, several open problems and challenges need to be addressed. One of the key challenges is energy optimization. Optimally managing and distributing the available energy resources among the sensor devices and MCVs to ensure uninterrupted and efficient charging is a complex task. Developing energy optimization algorithms that consider the dynamic nature of the network, the mobility of the MCVs, and the varying energy demands of the devices is crucial for maximizing energy utilization and network performance.

Scalability is another important challenge in deploying WRSNs with multiple MCVs. As the network expands to accommodate a larger number of sensor devices and MCVs, maintaining efficient communication, coordination, and resource allocation becomes increasingly complex. Designing scalable protocols and mechanisms that can handle the increased network size while minimizing overhead and preserving energy efficiency is essential for the widespread adoption of ISAC-assisted WRSNs. Additionally, cost-effectiveness, interoperability, and security also pose significant challenges that require further research and innovation to ensure the successful implementation and operation of these charging solutions in real-world IoT applications.

As the network expands to accommodate a larger number of sensor devices and MCVs, maintaining efficient communication, coordination, and resource allocation becomes increasingly complex.

References

[1] A. Khan *et al.*, "Multilevel Privacy Controlling Scheme to Protect Behavior Pattern in Smart IoT Environment," *Wireless Commun. and Mobile Computing*, vol. 2021, 2021, pp. 1–17.
[2] M. U. F. Qaisar *et al.*, "SDORP: SDN Based Opportunistic Routing for Asynchronous Wireless Sensor Networks," *IEEE Trans. Mobile Computing*, 2022.
[3] F. T. Wedaj *et al.*, "Reco: On-Demand Recharging and Data Collection for Wireless Rechargeable Sensor Networks," *IEEE Trans. Green Commun. and Net.*, 2023.
[4] A. Kaswan, P. K. Jana, and S. K. Das, "A Survey on Mobile Charging Techniques in Wireless Rechargeable Sensor Networks," *IEEE Commun. Surveys & Tutorials*, vol. 24, no. 3, 2022, pp. 1750–79.
[5] W. Yuan *et al.*, "Integrated Sensing and Communication-Assisted Orthogonal Time Frequency Space Transmission for Vehicular Networks," *IEEE J. Selected Topics in Signal Processing*, vol. 15, no. 6, 2021, pp. 1515–28.
[6] Q. Qi *et al.*, "Integrating Sensing, Computing, and Communication in 6G Wireless Networks: Design and Optimization," IEEE *Trans. Commun.*, vol. 70, no. 9, 2022, pp. 6212–27.
[7] T. Rault, "Avoiding Radiation of On-Demand Multi-Node Energy Charging with Multiple Mobile Chargers," *Computer Commun.*, vol. 134, 2019, pp. 42–51.
[8] W. Xu *et al.*, "Minimizing the Maximum Charging Delay of Multiple Mobile Chargers Under the Multinode Energy Charging Scheme," *IEEE Trans. Mobile Computing*, vol. 20, no. 5, 2020, pp. 1846–61.
[9] A. Kaswan *et al.*, "Dmcp: A Distributed Mobile Charging

Protocol in Wireless Rechargeable Sensor Networks," *ACM Trans. Sensor Networks*, vol. 19, no. 1, 2022, pp. 1–29.
[10] L. Mo, A. Kritikakou, and S. He, "Energy-Aware Multiple Mobile Chargers Coordination for Wireless Rechargeable Sensor Networks," *IEEE Internet of Things J.*, vol. 6, no. 5, 2019, pp. 8202–14.
[11] G. Han *et al.*, "An Uneven Cluster-Based Mobile Charging Algorithm for Wireless Rechargeable Sensor Networks," *IEEE Systems J.*, vol. 13, no. 4, 2018, pp. 3747–58.
[12] A. Tomar, L. Muduli, and P. K. Jana, "A Fuzzy Logic-Based On-Demand Charging Algorithm for Wireless Rechargeable Sensor Networks with Multiple Chargers," *IEEE Trans. Mobile Computing*, vol. 20, no. 9, 2020, pp. 2715–27.
[13] Y. Zhu *et al.*, "Velocity Control of Multiple Mobile Chargers Over Moving Trajectories in RF Energy Harvesting Wireless Sensor Networks," *IEEE Trans. Vehic. Tech.*, vol. 67, no. 11, 2018, pp. 11,314–18.
[14] M. U. F. Qaisar *et al.*, "Probabilistic On-Demand Charging Scheduling for ISAC-Assisted WRSNs with Multiple Mobile Charging Vehicles," *IEEE GLOBECOM 2023*, 2023, pp. 5895–5900.
[15] M. U. F. Qaisar *et al.*, "Poised: Probabilistic On-Demand Charging Scheduling for ISAC-Assisted Wrsns with Multiple Mobile Charging Vehicles," *IEEE Trans. Mobile Computing*, 2024.

Biographies

Muhammad Umar Farooq Qaisar [M] (muhammad@sustech.edu.cn) received his B.S. degree from the International Islamic University, Islamabad, Pakistan, in 2012, and his M.S. degree in Computer Science and Technology from the University of Science and Technology of China in 2017. He earned his Ph.D. degree in Computer Science and Technology from the University of Science and Technology of China in 2022. He was a postdoctoral fellow with the School of System Design and Intelligent Manufacturing at the Southern University of Science and Technology. Currently, he is an Associate Professor in the School of Computer Science at Northwestern Polytechnical University. His main research interests include IoT, WSN, SDN, VANETs, ISAC, UAVs, and communication security.

Weijie Yuan [M] (yuanwj@sustech.edu.cn) received the B.E. degree from the Beijing Institute of Technology, China, in 2013, and the Ph.D. degree from the University of Technology Sydney, Australia, in 2019. In 2016, he was a Visiting Ph.D. Student with the Institute of Telecommunications, Vienna University of Technology, Austria. He was a Research Assistant with the University of Sydney, a Visiting Associate Fellow with the University of Wollongong, and a Visiting Fellow with the University of Southampton, from 2017 to 2019. From 2019 to 2021, he was a Research Associate with the University of New South Wales. He is currently an Assistant Professor with the Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China. He was a recipient of the Best Ph.D. Thesis Award from the Chinese Institute of Electronics and an Exemplary Reviewer from IEEE TCOM/WCL. He currently serves as an Associate Editor for the IEEE Communications Letters, an Associate Editor and an Award Committee Member for the EURASIP Journal on Advances in Signal Processing. He has led the guest editorial teams for three special issues in *IEEE Communications Magazine*, *IEEE Transactions on Green Communications and Networking*, and *China Communications*. He was an Organizer/the Chair of several workshops and special sessions on orthogonal time frequency space (OTFS) and integrated sensing and communication (ISAC) in flagship IEEE and ACM conferences, including IEEE ICC, IEEE GLOBECOM, IEEE/CIC ICCC, IEEE SPAWC, IEEE VTC, IEEE WCNC, IEEE ICASSP, and ACM MobiCom. He is the Founding Chair of the IEEE ComSoc Special Interest Group on Orthogonal Time Frequency Space (OTFS-SIG).

Paolo Bellavista [SM] (paolo.bellavista@unibo.it) received the Ph.D. degree in computer science engineering from the University of Bologna, Italy, in 2001. He is currently a Full Professor with the University of Bologna. His research interests include middleware for mobile computing, QoS management in the cloud continuum, infrastructures for big data processing in industrial environments, and performance optimization in wide-scale and latency-sensitive deployment environments. He serves on the Editorial Boards of *IEEE Communications Surveys and Tutorials*, *IEEE Transactions on Network and Service Management*, *IEEE Transactions on Services Computing*, *ACM CSUR*, *ACM TIOT*, and *PMC* (Elsevier). He recently served as the Scientific Coordinator of the H2020 IoTwins Project (https://www.iotwins.eu).

Guangjie Han [F] (hanguangjie@gmail.com) is currently a Professor with the Department of Internet of Things Engineering, Hohai University, Changzhou, China. He received his Ph.D. degree from Northeastern University, Shenyang, China, in 2004. In February 2008, he finished his work as a Postdoctoral Researcher with the Department of Computer Science, Chonnam National University, Gwangju, Korea. From October 2010 to October 2011, he was a Visiting Research Scholar with Osaka University, Suita, Japan. From January 2017 to February 2017, he was a Visiting Professor with City University of Hong Kong, China. From July 2017 to July 2020, he was a Distinguished Professor with Dalian University of Technology, China. His current research interests include Internet of Things, Industrial Internet, Machine Learning and Artificial Intelligence, Mobile Computing, Security and Privacy. He has over 500 peer-reviewed journal and conference papers, in addition to 160 granted and pending patents. Currently, his H-index is 66 and i10-index is 292 in Google Citation (Google Scholar). The total citation count of his papers raises above 16500+ times. He is a Fellow of the UK Institution of Engineering and Technology (FIET). He has served on the Editorial Boards of up to 10 international journals, including the IEEE TII, IEEE TCCN, IEEE TVT, IEEE Systems, etc. He has guest-edited several special issues in IEEE Journals and Magazines, including the *IEEE JSAC*, *IEEE Communications*, *IEEE Wireless Communications*, *Computer Networks*, etc. He has also served as chair of organizing and technical committees in many international conferences. He has been awarded 2020 IEEE Systems Journal Annual Best Paper Award and the 2017–2019 IEEE ACCESS Outstanding Associate Editor Award.

Adeel Ahmed [StM] (adeelahmed@mail.ustc.edu.cn) received the BS degree in Telecommunication and Networking from COMSATS University, Pakistan in 2015 and received his M.S. degree in Computer Science and Technology from University of Science and Technology of China in July 2021. Currently, he is pursuing Ph.D. in Computer Science and Technology from University of Science and Technology of China. His research interests include IoT, WBAN, SDN, and SDR.

IEEE World Forum on Internet of Things

IEEE Internet of Things

The 10th IEEE World Forum on the Internet of Things (IoT)

10–13 November 2024 // Ottawa, Canada

An In Person Event

Unleashing the Power of IoT with AI

wfiot2024.iot.ieee.org

CALL FOR PAPERS

The 10th anniversary of the World Forum on the Internet of Things (WF-IoT2024) marks a significant milestone for this distinguished IEEE IoT conference and gathering. Recognized as a premier event in the IoT landscape, WF-IoT serves as a vital platform for both academic and industry professionals passionate about Internet of Things (IoT) advancements. WF-IoT seeks submissions and proposals for original technical papers and presentations that address the Internet of Things (IoT), its theoretical and technological building blocks, the applications that drive the growth and evolution of IoT, operational considerations, experimentation, experiences from deployments, and the impacts of IoT on consumers, the research community, the public sector, as well as commercial and industrial sectors. This anniversary edition promises to uphold the tradition of fostering collaboration, knowledge exchange, and exploration of the latest trends and innovations within the dynamic realm of the Internet of Things. The theme for WF-IoT2024 is "Unleashing the power of IoT with AI", encouraging the submissions of content focused on aspects of IoT that cover most of our planet and the near-space environment. The World Forum will be held in-person.

All accepted Peer-Reviewed Technical Papers will be included in the proceedings (in IEEE Xplore). A complete list of topics and directions can be found on the IEEE World Forum on the Internet of Things Website: **https://wfiot2024.iot.ieee.org/**

IMPORTANT DATES:

Full Paper Submission Date:	**1 April 2024**
Notification of Acceptance Date:	**15 July 2024**
Final Paper Submission Date:	**1 August 2024**
Submission Link:	**https://edas.info/newPaper.php?c=31878**

SPECIAL SESSIONS, WORKSHOPS, AND INDUSTRY FORUMS

As part of the Technical Program, the World Forum is also seeking proposals for (1) Special Sessions consisting of peer-reviewed papers focused on research topics of importance to IoT; (2) Workshops, consisting of peer-reviewed papers, discussions, keynotes and summary results about advanced topics relevant to IoT. Workshop summary results will be edited and published as part of the WF-IoT2024 Proceedings; and (3) Industry Forums consisting of presentations, executive forums and panel discussions aimed at research topics important to industrial IoT issues. The proposals are due by 20 April 2024. Once accepted, proposers for Special Sessions and Workshops will issue an individual call for papers for selected topics.

Paper submissions for Special Sessions and Workshops are due 15 May 2024.

ADDITIONAL PARTS OF THE WF-IOT2024 PROGRAM

WF-IoT2024 will also feature the following:

- Plenary Sessions
- Tutorials on Research and Industrial Hot Topics
- Women in Engineering (WIE) Forum
- Young and Professionals (YP) Forum
- Masters/PhD Forum and Student Paper Contest (MP)
- Entrepreneurial Forum

The deadline for speaker nominations and content for the WIE, YP, and Entrepreneurial Program is 1 June 2024.

IEEE ComSoc · CEDA · CAS · IEEE Reliability Society · IEEE Signal Processing Society · IEEE CSS · Electron Devices Society · IEEE Industrial Electronics Society · MTT-S · IEEE PES · pels · IEEE Solid-State Circuits Society

Unleashing the Potential of Aerial RISs in Post-Disaster Scenarios

Maurilio Matracia, Mustafa A. Kishk, and Mohamed-Slim Alouini

Abstract

Conventional wireless network infrastructures are known to be susceptible to strong perturbations such as the ones caused by calamities. With this regard, whenever facing an emergency, it is vital to ensure reliable connectivity within the disaster-struck zone. Therefore, in this article we promote the use of aerial reconfigurable intelligent surfaces (RISs) as a possible solution for supporting any possible damages affecting terrestrial base stations (TBSs). At the same time, we discuss the main differences between the aerial RIS (ARIS) technology and its parent ones, namely the terrestrial RIS (TRIS) and the aerial base station (ABS). To support our vision, we recall hurricane Maria, which affected Puerto Rico in 2017, and propose insightful real-world-inspired simulation results in order to discuss what we believe are the main challenges for a commercial implementation of ARISs in post-disaster scenarios.

Introduction

Disasters have always represented and still represent a threat to communities, since they are capable of impairing life, economy, and activities within the suffered region. In fact, the United Nations (UNs) decided to establish the International Day for Disaster Risk Reduction (IDDRR), on October 13th, to stimulate all populations and governments in improving their resilience to calamities. This is particularly important nowadays as the number of world's severe earthquakes (of at least magnitude 7.0 in the *modified Mercalli intensity scale*) that happened in 2022 is just twice compared to the first two months of 2023. For example, during the writing phase of this manuscript a tremendous 7.8 magnitude earthquake affected Turkey and Syria, killing over forty thousand people.

One critical phase of the disaster management process regards the development of effective search and rescue (SAR) procedures, which are mostly based on human, canine, or electronic agents. Although humans do not necessarily require to be professionals rescuers, their participation exposes them to huge risks. While the advantages of canine missions are particularly important in case of unconscious trapped victims, even well-trained dogs cannot be supportive when the amount of rubble is excessive. Finally, rescuing the victims by means of electronic equipment leverages acoustic and vibration signals produced by conscious casualties [1].

In any case, the use of wireless communication technologies is indispensable for enhancing the chances of saving lives and minimizing the losses caused by the calamity. Although wireless communication systems are often given for granted when it comes to provide daily services, their availability usually becomes insufficient in emergency situations; in fact, the failures of some network infrastructure's parts and the increased service demand could damage, overload, and even isolate them.

The common counteraction in the post-disaster phase is to deploy ad hoc networks; in particular, the ones based on the deployment of unmanned aerial vehicles (UAVs) such as drones or high-altitude platforms (HAPs)} are increasingly attracting the attention of academic and industrial communities. This solution can be implemented in various ways, but the most interesting from our perspective are the ones where the UAV carries either a base station (BS) equipment or an RIS [2], which is a programmable surface structure capable of reflecting electromagnetic waves. Figure 1 provides an illustration of a HAP-mounted RIS for emergency situations. However, while UAV-mounted BSs are already available in the market (although not yet on a large scale), RISs are still at the prototyping stage [3]; therefore, we would like to emphasize that the discussions about these advanced reflectors (either in their aerial or terrestrial fashions) are based on just a few preliminary experimental results and not on a thorough in-field and commercial experience.

Contributions of this Article

Compared to the existing literature on post-disaster communications, we can summarize the main contributions of this article by highlighting our proposed:

- Extensive comparisons between the novel ARIS technology and its parents, namely the TRIS and ABS technologies.
- Real-world-inspired case study (based on detailed datasets) demonstrating the potential benefits of deploying an ARIS to support the surviving cell towers.
- Overview of the challenges and open problems regarding the use of ARISs.

Maurilio Matracia and Mohamed-Slim Alouini are with King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia (KSA); Mustafa Kishk is with Maynooth University, Maynooth, Ireland.

Digital Object Identifier: 10.1109/IOTM.001.2300258

2576-3180/24/$25.00 © 2024 IEEE

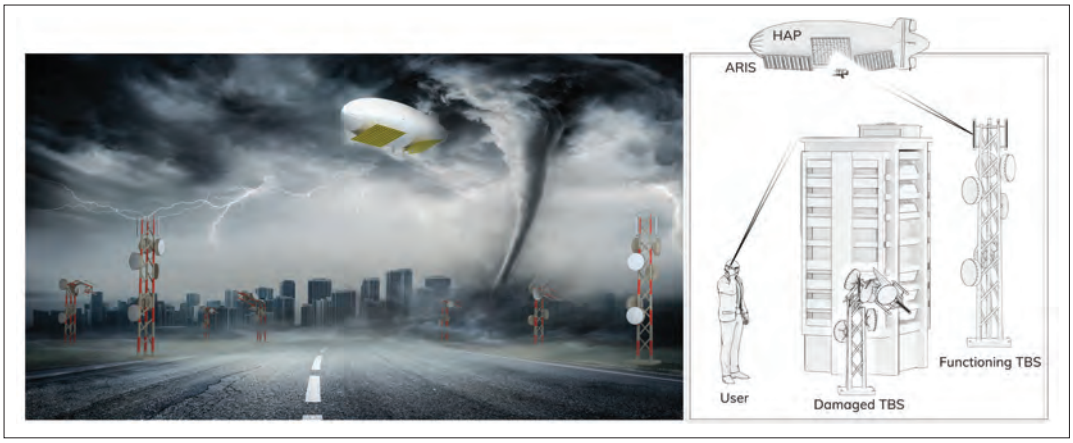


FIGURE 1. System figure: ARISs can be deployed in post-disaster scenarios to improve the quality of the communication links between users and surviving TBSs.

Outline of this Article

The remainder of this article is structured as follows:

- In the next section, we provide context on the existing literature works about ARISs and post-disaster communications (PDCs).
- Then, we discuss the main differences between the ARIS and TRIS paradigms, with a special focus on the deployment aspects.
- Similarly, we compare the ARIS technology with the ABS one in terms of hardware, power consumption, susceptibility to disturbances, and delay.
- The fifth and core section of the article provides insightful simulation results about the potential benefits of the ARIS' deployment in a disaster-prone environment such as the island of Puerto Rico. Our simulations leverage stochastic geometry (SG) tools as well as real-world data sets regarding the cell tower's and population density's distributions.
- We suggest some future directions of our research and enlists the main challenges regarding the use of ARISs (and especially drones) in post-disaster scenarios.
- Finally, we summarize and conclude this article.

Existing Works on ARISs for PDCs

Firstly, we should mention [4], where the authors promoted the integration of RISs and UAVs (either as RIS-equipped or RIS-assisted UAVs) to carry on public safety missions; however, the proposed study provided results about the achievable data rate by considering only the serving TBS, without taking into account any possible interferers.

In addition, authors in [1] designed the so-called *pseudo-multilateration*, a novel localization technique which, differently from the classical multilateration one, considers a single moving anchor (for example, a drone) to estimate the distance from a user over time; to this extent, the authors suggested deploying ARISs in order control the channel and find its optimal configuration.

Then, authors in [5] assumed a fixed-wing[1] UAV-mounted RIS to serve trapped UEs under Fisher-Snedecor-*F* composite fading conditions; the study lead to novel closed-form expressions for the bit error rate (BER), channel statistics, and outage probability.

Finally, our recent work [6] introduced an SG-based mathematical framework to evaluate the network performance in post-disaster scenarios. In particular, we derived the distance distributions, association probabilities, and Laplace transform of the interference in order to compute both the local and the average coverage probabilities in case of a circular disaster-struck area; then, we showed that a reasonably-designed ARIS-aided architecture generally outperforms its ABS-aided and non-aided counterparts.

However, compared to all the aforementioned works, the peculiarity of this article consists in providing results that are based on data sets extracted from real geographic areas.

Moving on, Fig. 2 illustrates a qualitative comparison between the ARIS technology and its parents (TRIS and ABS), while more detailed discussions will be provided in the next two sections.

Comparing ARISs and TRISs

This section provides a bird's eye view of the main differences between ARISs and TRISs in terms of channel conditions and deployment.

Line-of-Sight (LoS) Communications

In conventional terrestrial networks (with ground users and BSs), the region of influence of a TRIS is more limited than its ARIS counterpart due to substantial differences in terms of surface orientation. Indeed, TRISs are mostly mounted on buildings and therefore they are exposed to just half of the ground plane, hence the coverage region is very limited compared to its ARIS counterpart. However, for similar reasons the presence of multiple non-terrestrial relays (either aerial or space platforms) can strongly promote the TRIS over the ARIS, since the latter are generally facing downwards.

Finally, the high altitude makes the reflected signals less likely to be obstructed by buildings and trees, and therefore one single reflection is usually enough to reach the desired user (often implying negligible signal attenuation). Nonetheless, increasing the altitude also increases the distance to the ground plane, leading to a trade-off between a higher LoS probability and a higher path loss when evaluating the overall channel conditions.

Vibrations

Since a TRIS represents a static node, its vibrations are negligible and hence do not lead to considerable channel fluctuations. On the other side,

In conventional terrestrial networks (with ground users and BSs), the region of influence of a TRIS is more limited than its ARIS counterpart due to substantial differences in terms of surface orientation. Indeed, TRISs are mostly mounted on buildings and therefore they are exposed to just half of the ground plane, hence the coverage region is very limited compared to its ARIS counterpart.

[1] Despite the limited maneuverability, the authors decided to focus on the fixed-wing design because it is capable of carrying a larger RIS compared to an equivalent rotary-wing design

Deploying an ARIS (and hence, a UAV) may raise safety issues and privacy concerns (although the latter could be neglected in emergency scenarios).

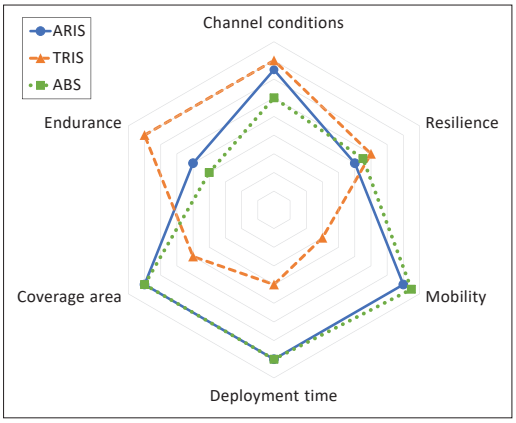


FIGURE 2. Radar chart: qualitative comparison between ARIS, TRIS, and ABS technologies.

the missions for UAVs, and especially untethered drones, are characterized by persistent vibrations; although using multiple ropes to carry the RIS can help in mitigating the wobbles, these may still be not sufficient for achieving high accuracy in terms of channel estimation and beam steering [7, Sec. II-A].

Deployment

Generally speaking, deploying a TRIS is much more time-consuming than deploying an ARIS. Indeed, for the TRIS case, the deployment phase starts from the choice of a suitable site (usually a building facade), and thus it implies complications due to its visual impact and the availability of the building's owner[2] [7, Sec. II-A]. In a post-disaster scenario, however, deploying a TRIS on buildings would not often be feasible due to time constraints as well as the lack of a proper personnel, but deploying it on an ad hoc ground vehicle would still be an option.

Deploying an ARIS (and hence, a UAV) may raise safety issues and privacy concerns (although the latter could be neglected in emergency scenarios). In addition, technological limitations in terms of the UAV's endurance and susceptibility to harsh weather conditions need to be taken into account before starting any mission; such constraints are much more relaxed for the case of a TRIS, since it is supported by a fixed structure. However, some TRIS could have been deployed even before the occurrence of a disaster, hence the fragility of the RIS and the limited resilience of the structure hosting it may represent a critical factor (especially in case of strong earthquakes). Finally, while the TRIS is typically fixed, the ARIS enjoys much higher mobility and relocation flexibility [8].

Comparing ARISs and ABSs

In this section, we overview the main differences between ARISs and ABSs under various important aspects.

Hardware

ABSs include several electronic elements (e.g., digital-to-analog and analog-to-digital converters, mixers, and amplifiers for transmission and reception), and the number significantly increases in case of full-duplex (FD) relaying, which evidently leads to relatively high computational costs [7, Sec. II-A]. On the other side, the metallic or dielectric nature of the RISs' patches (properly combined with low-power active components such as switches and varactors) leads to high configurability without any problem related to antenna noise amplification nor self-interference [8]; hence, no complicated electronic circuits and numerous active components are required by ARISs [9], leading also to a smaller probability of failure and hence a higher resilience compared to ABSs. In other words, we can consider the ARIS as a simpler (yet effective) amplify-and-forward (AF) ABS.

Finally, while the BS equipment can be attached to the frame of a drone, if the same vehicle carries an RIS by means of ropes it is more unstable from an aerodynamical point of view, and hence more likely to accidentally collide with other objects; this endows ABSs of a slightly higher resilience and mobility compared to ARISs.

Power Consumption

Power consumption is a critical aspect for untethered UAVs, and especially drones. While relays are generally equipped with all the aforementioned active components (which consume a considerable amount of power, apart from the power consumed during radio frequency transmission), the RISs' passive array architecture leads to the same functionalities of large antenna arrays at a fraction of the energy consumption; the only power required by the RIS is to feed its control unit [4, Sec. II-A], and hence ARISs are expected to achieve slightly longer endurance compared to equivalent ABSs. However, there is an exception when just a low rate is targeted, and in that case the decode-and-forward (DF) relay would be more energy-efficient [4, Sec. 3.2].

As a reference, the prototype recently introduced in [3] achieved a power consumption as small as 1 W for a board of 1100 elements clustered in roughly one-fourth of a meter square.

Disturbances

ABSs can operate according to various relaying protocols, but one important aspect to take into account is the presence of disturbances such as additive noise and loop-back self-interference. Whenever adopting the AF relaying protocol, the performance of ABSs may be compromised due to additive noise. On the other side, the DF relaying protocol avoids this issue, but requires decoding and re-encoding the signal (which leads to higher computational and energy costs). Furthermore, in case of FD operation mode the effect of residual loop-back self-interference negatively affects the quality of the communication [9].

RISs, instead, are immune from additive noise (but not from phase noises). Finally, when considering a large scale network, it is fair to assume that the RIS will not suffer the interference coming from every TBS, but will be able to detect the signal coming from the desired direction; since a similar assumption cannot be made for the ABS case (unless advanced massive multiple-input multiple-output (MIMO) techniques are implemented at the cost of higher computational powers as well as heavier and bulkier equipment), we conclude by saying that ARISs enjoy better channel conditions than ABSs.

Transmission Delay

[2] However, some forms of transparent RIS are currently being developed in order to solve these problems.

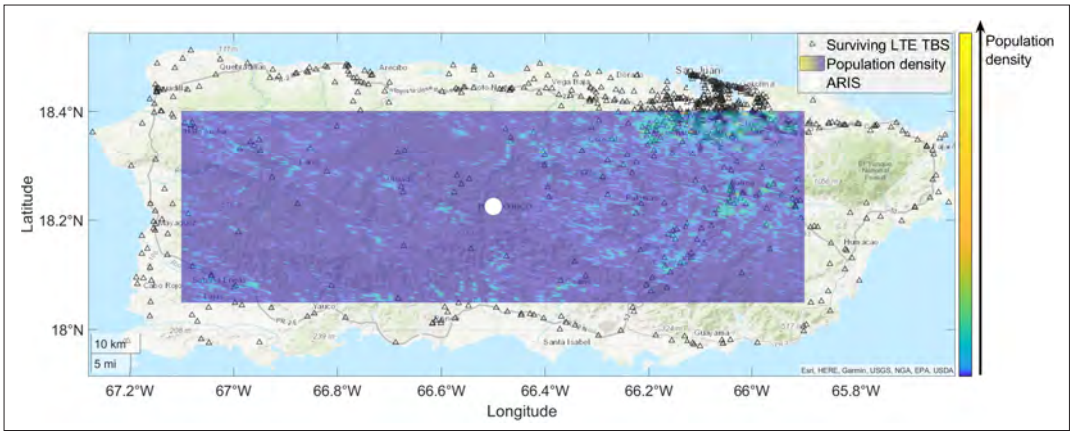


FIGURE 3. System setup: the ARIS is deployed at the center of a rectangular area of interest (within which the population density is taken into account) to support the surviving cell towers (the damaged ones are omitted).

While FD relaying suffers from high energy and computational costs as well as loop-back self-interference, its advantage is to require only one time slot to transmit a signal. Hence, the transmission delay of an ABS operating in FD mode is comparable to the one of an ARIS; however, a common ABS operates in half-duplex (HD) mode, and suffers from a transmission delay that lasts roughly twice longer [7, Sec. II-A].

Real-World Case Study

System Setup

For our study we picked Puerto Rico, an island of the Atlantic Ocean suffering frequent exposure to storms and floods. In particular, Puerto Rico has become very famous among the telecom community because of the hurricane Maria and the consequent ad hoc deployment of *Project Loon's* balloons [11], which was able to support around a hundred thousand users.

In this occasion the existing cell sites' quality of resilience (QoR) was as little as 5%, meaning that only one TBS every twenty survived the calamity. An illustration of this situation is shown by Fig. 3, where 95% of the original TBSs (extracted from the *OpenCellid* data set [12]) have been randomly removed, yet resulting in a considerable overall number.

However, the serving TBS will be selected among a subset of the surviving ones, since we assume that a LoS link (either direct or indirect) is needed by both the user and the ARIS in order to properly detect the signal; therefore, for each possible link we will determine if its ends are in LoS or non-LoS (NLoS) condition by following the stochastic approach presented in [10, Sec. II]. The details about our assumptions on the channel conditions of the ARIS and the ABS are hereby omitted, but can be extracted from our technical paper [6].

Finally, for our study we will consider a rectangular area, within which the population density's distribution (as provided by the data set of the Humanitarian Data Exchange (HDX) [13]) is taken into account when selecting the typical user. The ARIS is assumed to hover above the center of the area of interest.

Results and Discussion

Our goal is to quantify the improvement, in terms of coverage probability, that the deployment of an HAP-mounted RIS can bring in the post-disaster scenarios. To this extent, we partitioned the area of interest in multiple rectangular sub-areas, and computed the coverage probability (averaged over a large number of iterations and weighted on the population density) over each sub-area.

Compared for example with the relay technology, the main advantage of the RIS is that it can be fairly assumed capable of accurately selecting the signals that come from a specific direction and precisely reflecting them along another direction; in other words, the RIS would not cause any interference while the relay would, unless advanced massive MIMO techniques (which are not always compatible with UAVs) are used. Our results have been obtained via Monte Carlo simulations by applying the following procedure:

- One user per sub-area is selected based on the population density distribution.
- For each user, the maximum average received power association rule is applied by comparing the power coming from the closest LoS TBS with the power received through the best indirect LoS link.
- The resulting SINR is computed (by taking into account that only direct links provide interference) and compared with the respective threshold.
- The previous points are repeated at each iteration and the coverage probability is computed for each sub-area.

The simulation parameters are listed in Table 1. Note that while $M = 10^6$ may seem excessive for an RIS, by scaling the prototype in [3] and considering the weight of each element equal to 10 g as in [14], this would respectively correspond to an area of almost $250M^2$ and an overall weight of around 10^4 kg, which is compatible with both size and payload capability of a large blimp. Finally, we set an altitude of 50 km as commonly done in the literature when seeking wide coverage areas.

Figure 4a and Fig. 4b clearly show that properly deploying an ARIS can be vital not only for the users affected by the calamity, but also for the first responders. In particular, we can see massive improvements in the North-West and South-East zones of the area of interest, where the deployment of the ARIS leads to an improvement of around 10–15%; on the other hand, either with or without the ARIS the coverage probability

While FD relaying suffers from high energy and computational costs as well as loop-back self-interference, its advantage is to require only one time slot to transmit a signal.

Parameter	Value
Number of Monte Carlo iterations	$n_i = 10^4$
TBSs' QoR	$\chi = 5\%$
TBSs' altitude	$h_T = 30$ m
TBSs' transmit power	$p_T = 20$ W
ARIS' altitude	$h_A = 50$ km
ARIS' number of elements	$M = 10^6$
Noise power	$N_0 = 10^{-11}$, W
SINR threshold	$\tau = 0.3162 = -5$ dB
LoS Nakagami-*m* shape parameter	$m_L = 3$
LoS path loss exponent	$a_L = 2.3$
S-curve parameters [10]	$a = 4.88$; $b = 0.429$

TABLE 1. Main simulation parameters.

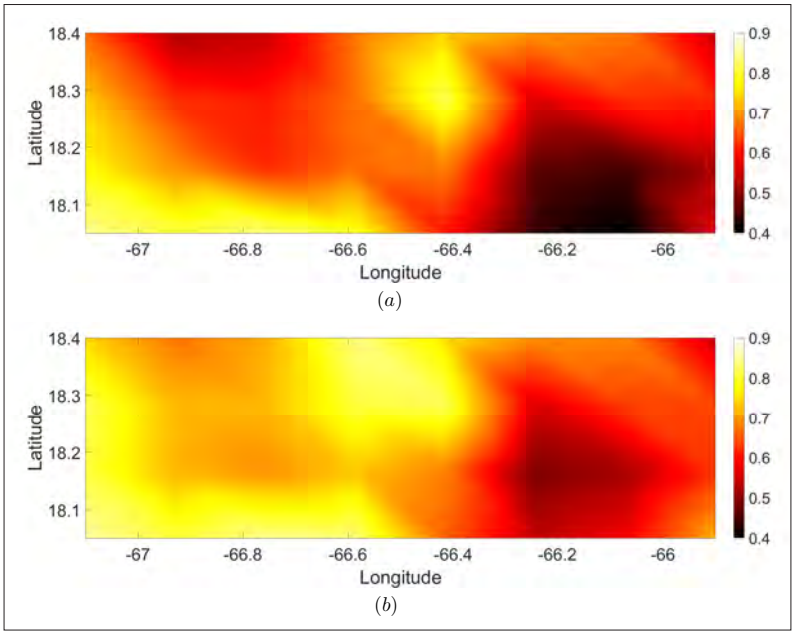


FIGURE 6. Coverage probability over the considered area in Puerto Rico when: a) no ARIS is deployed; and b a HAP-mounted RIS with one million antenna elements is deployed at an altitude of 50km above the center of the same area.

remains fixed at around 60% in the North-East corner, because of the relatively high density of TBSs. In conclusion, by taking the average of the performance over all the sub-areas, the network's coverage jumped from 63.9% to 70.2% when deploying the HAP-mounted RIS (while the average coverage in case of an equivalent ABS deployment would be just 67.4%). Given the extension of the area of interest and the fact that just one RIS was deployed, such a considerable boost was achieved thanks to both the high altitude h_A and the large number of elements M, which allowed the signals coming from the TBSs to respectively take a path with less obstacles and be amplified.

Challenges and Future Directions

Challenges

Whenever considering aerial vehicles, both their endurance and resilience to harsh weather represent challenging aspects; in particular, UAVs would strongly suffer in case of prolonged winds and tornadoes, while for example the presence of smoke would affect all wireless communication systems in similar ways.

Moreover, the use of RISs in case of terrorist attacks brings important security concerns due to the fact that they are very sensitive to the directions of the signals (as already mentioned earlier): as an example, any malicious UAV could be placed between the RIS and the serving TBS (or the user) in order to transmit fake information (or eavesdrop the one intended for the user).

Finally, accurate modeling of the channel from/to a drone-mounted RIS is still very complicated due to its continuous wobbling, especially if it operates at high-frequency signals; moreover, given the drone's mobility, channel modeling would become even more challenging if the drone does not inform its serving TBS about its trajectory plan. Therefore, at least for the cases where the reflectors are carried by means of ropes, we believe in the importance of designing aerodynamic and flexible RIS frames to avoid accumulation of rain on the reflector, as well as turbulence and collisions.

Future Directions

Following the same direction of this case study, it would be interesting to investigate any possible integration of the ARIS with the closest satellite systems; in particular, by taking into account the expected trajectories of the platforms, the ARIS' location in the tridimensional space (at least for the case UAVs endowed with propellers).

In addition, the system setup may include a swarm of ARISs rather than just a single one, especially in case of relatively-small disaster areas (as we already demonstrated for ABS-aided PDCs [15]). For instance, in some cases deploying multiple drones may be a better solution compared to deploying just one single HAP: indeed, a large fleet would not only lead to a better capillarity of the services (due to both the shorter path losses and higher probabilities of seeing the users), but would also bring the opportunity of easily serving more users at the same time.

Summary and Conclusion

In this article, we have extensively discussed the deployment of ARISs for PDCs, with a special focus on their differences with ABSs and TRISs; as far as we are concerned, in public safety scenarios the main advantage of aerial nodes lies in their short deployment time, and combining this with a technology that does not provide interference (such the RIS) would make emergency communications much more effective.

In fact, even without considering any dependence on time, our results showed the coverage gain that a HAP-mounted RIS could bring to a Puerto Rican area of several thousands of square kilometers is considerable. However, it is evident that as of today such solution is also subject to several limitations, for instance in terms of controllability in case of strong winds.

Nonetheless, we hope that the huge research interest in this topic will be soon justified by conspicuous governments' investments, considerable technological advancements, and spread of commercial products on a global scale.

Acknowledgements

The authors would like to acknowledge that Fig. 1 was produced by Ana Bigio, scientific illustrator.

References

[1] A. Albanese, V. Sciancalepore, and X. Costa-Pérez, "First Responders Got Wings: UAVs to the Rescue of Localization Operations in Beyond 5G Systems," *IEEE Commun. Mag.*, vol. 59, no. 11, 2021, pp. 28–34.

[2] S. Alfattani *et al.*, "Aerial Platforms with Reconfigurable Smart Surfaces For 5G and Beyond," *IEEE Commun. Mag.*, vol. 59, no. 1, 2021, pp. 96–102.

[3] X. Pei *et al.*, "RIS-Aided Wireless Communications: Prototyping, Adaptive Beamforming, and Indoor/Outdoor Field Trials," *IEEE Trans. Commun.*, vol. 69, no. 12, 2021, pp. 8627–40.

[4] W. Jaafar *et al.*, "Enhancing UAV-Based Public Safety Networks with Reconfigurable Intelligent Surfaces," *Intelligent Unmanned Air Vehicles Communications for Public Safety Networks*, Springer, 2022, pp. 145–67.

[5] Y. Chen and W. Cheng, "Performance Analysis of RIS-Equipped-UAV Based Emergency Wireless Communications," *IEEE Int'l. Conf. Commun. (ICC)*, Seoul, South Korea, 2022, pp. 255–60.

[6] M. Matracia, M. A. Kishk, and M.-S. Alouini, "Comparing Aerial-RIS- and Aerial-Base-Station-Aided Post-Disaster Cellular Networks," *IEEE Open J. Vehic. Tech.*, 2023, pp. 1–15.

[7] J. Ye *et al.*, "Non-Terrestrial Communications Assisted by Reconfigurable Intelligent Surfaces," *Proc. IEEE*, 2022.

[8] H. Lu *et al.*, "Enabling panoramic full-angle reflection via aerial intelligent reflecting surface," *IEEE Int'l. Conf. Commun. Wksps. (ICC Workshops)*, Dublin, Ireland, 2020, pp. 1–6.

[9] M. Di Renzo *et al.*, "Reconfigurable Intelligent Surfaces vs. Relaying: Differences, Similarities, and Performance Comparison," *IEEE Open J. Commun. Society*, vol. 1, 2020, pp. 798–807.

[10] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP Altitude for Maximum Coverage," *IEEE Wireless Commun. Letters*, vol. 3, no. 6, 2014, pp. 569–72.

[11] D. Lumb, "Project Loon Delivers Internet to 100,000 People in Puerto Rico," 2017. Engadget, accessed on, Feb. 5, 2023; https://www.engadget.com/2017-11-09-project-loon-delivers-internet-100-000-people-puerto-rico.html.

[12] Unwired Labs, "310," 2023, accessed on Feb. 5, 2023; https://opencellid.org/downloads.php?token=pk.15b372a091936dc3e966d46c5b1315a1.

[13] Humanitarian Data Exchange, "population pri 2018-10-01.csv.zip," 201, accessed on Feb. 5, 2023; https://data.humdata.org/dataset?res format=CSV&organization=facebook&q=puerto%20rico&sort=if(gt(last modified%2Creview date)%2Clast modified%2Creview date)%20desc&ext page size=25.

[14] D. Tyrovolas *et al.*, "Energy-Aware Design of UAVmounted RIS Networks for IoT Data Collection," *IEEE Trans. Commun.*, vol. 71, no. 2, 2023, pp. 1168–78.

[15] M. Matracia, M. A. Kishk, and M.-S. Alouini, "On the topological aspects of UAV-Assisted Post-Disaster Wireless Communication Networks," *IEEE Commun. Mag.*, vol. 59, no. 11, 2021, pp. 59–64.

Biographies

Maurilio Matracia [S'21] (maurilio.matracia@kaust.edu.sa) is currently a Doctoral Student of the Communication Theory Lab (CTL) at KAUST. His experience with IEEE includes serving as a reviewer for several journals and receiving prizes at the *SusTech 2021 student poster*} as well as *ComSoc EMEA Region – Internet for All* contests. His main research interest is stochastic geometry, with a special focus on post-disaster and rural cellular networks.

Mustafa A. Kishk [S'16, M'18] (mustafa.kishk@mu.ie) received the B.Sc. and M.Sc. degrees from Cairo University, Giza, Egypt, in 2013 and 2015, respectively, and the Ph.D. degree from Virginia Tech, Blacksburg, VA, USA, in 2018, all in Electrical Engineering. He is an assistant professor at the Electronic Engineering Department, Maynooth University, Ireland. Before that, he was a Postdoctoral Research Fellow with the Communication Theory Laboratory, King Abdullah University of Science and Technology, Saudi Arabia. He currently serves as an associate editor with IEEE Wireless Communication Letters. His current research interests include stochastic geometry, UAV-enabled communication systems, and satellite-enabled communications. He is a recipient of the IEEE ComSoc Outstanding Young Researcher Award for Europe, Middle East, and Africa Region, in 2022. He was recognized as an Exemplary Reviewer by the IEEE Communications Letters in 2020 and 2021.

Mohamed-Slim Alouini [S'94, M'98, SM'03, F'09] (slim.alouini@kaust.edu.sa) was born in Tunis, Tunisia. He received his Ph.D. degree in Electrical Engineering from California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He is currently a Distinguished Professor of Electrical and Computer Engineering at KAUST. His current research interests include the modeling, design, and performance analysis of wireless communication systems.

Machine Learning-Based Predictive Inventory for a Vending Machine Warehouse

Umair Mehmood, John Broderick, Simon Davies, Ali Kashif Bashir, and Khaled Rabie

Abstract

In this study, we predict inventory for an IoT-enabled vending machine warehouse servicing approximately 1,500 vending machines with the goal of timely replenishing, achieving cost effectiveness, reducing stock waste, optimising the available resources and ensuring fulfilment of consumer demand. The study deploys four different ML algorithms, namely, Extreme gradient boosting, Autoregressive integrated moving average with/without exogenous variables (ARIMA/ARIMAX), Facebook Prophet (Fb Prophet), and Support Vector Regression (SVR). The study unfolds in two phases. First, we utilise conventional historical sales data variables to make the prediction whereas in the second phase, we systematically introduced external variables including weekday, sales deviation flag, and holiday flags into our ML algorithms. The results indicate a significant performance boost using external variables with extreme gradient boosting achieving the lowest (Mean Absolute Error) MAE of 22, followed by ARIMAX, FB Prophet, and SVR with MAE values of 27, 37, and 38, respectively.

Introduction

The transformative impact of internet of things (IoT) is not limited to a specific industry; it spans logistics, energy, healthcare, and, very recently, the vending industry. In a fast-paced era where consumers value convenience, vending machines have emerged as a pioneering solution to meet the ever-increasing demand for quick and accessible on-the-go snacking options, including crisps, chocolates and beverages. The key to successful vending operations lies in the ability to predict and manage vending machine inventory effectively to meet consumer demand. Accurate demand forecasting is essential for optimising revenue, profit margins, budgeting, risk mitigation, sales strategies, and capital expenditure [1].

The traditional vending machine restocking method involves manual stock monitoring by operators who then place orders to the warehouse for replenishment. This process is time-consuming, prone to human error and lacks real-time data for monitoring demand fluctuations, resulting in inaccurate forecasting and issues such as over or under-stocking. The overstocking of products leads to stock wastage, poor warehouse space management and blocked investment whereas understocking leads to lost sales opportunities and customer dissatisfaction.

To address these challenges, innovative methods utilising IoT devices have emerged to cater to the problem of monitoring inventory in vending machines. This method involves installing a telemetry unit in the vending machines which systematically capture the transactional data in real-time. Consequently, this provides insights into the number of items sold from the vending machine and the quantity required to replenish it to its maximum capacity in real-time. Using this approach yields several advantages such as eliminating the need of manually counting the required stock to order for the warehouse which is prone to human errors and improving the demand forecast since it is based on real-time data. However, this approach still does not capture the full spectrum of factors influencing inventory demand. It lacks the capability to account for seasonal variations and demand fluctuations linked to different days of the week, special events, holidays, weather conditions, etc. This limitation highlights the need for a more sophisticated solution, and this is precisely where the integration of machine learning (ML) comes into play. Therefore, within the scope of this study, we will be utilising ML algorithms to forecast vending machine warehouse inventory to fulfil the vending machine stock requirement ultimately enhancing operational efficiency reducing stock wastage, and, consequently, elevating revenue and customer satisfaction. For this study, we have chosen a real work scenario within the vending industry in the UK.

The remainder of this article is organised as follows. We present a review of the related works. We present the proposed system and data architecture as well as the use case considered in this study. We explore the data methodology and ML algorithms utilised. We devote a section to discussing the results and analysis whereas we then present the challenges and future research directions. Finally, we draw the main conclusions.

Background and Literature Review

Demand forecasting via ML has witnessed significant advancements in recent years and has been adopted by various industries to achieve cost effectiveness, optimise resource allocation and meet consumer demand. There are many different approaches that can be found in the literature utilised in various industries to solve the problem of demand forecasting. However, the implementation of these approaches in the vending industry has

This project is funded by Innovate UK (Grant number: 430309).

Digital Object Identifier: 10.1109/IOTM.001.2300271

Khaled Rabie and Ali Kashif Bashir are with Manchester Metropolitan University, UK.; Umair Mehmood, John Broderick, and Simon Davies are with Broderick Ltd, UK.

been very limited. Therefore, our research is a significant contribution to this field, as there is currently no existing literature examining demand forecasting within the context of vending machine networks.

The authors in [2], for instance, proposed demand forecasting of Stock Keeping Units (SKUs) by using RFID tags on them to monitor their movement in and out of the warehouse. This data is stored in real-time in the cloud and three ML models, namely, Support Vector Machine (SVM), K-nearest neighbour (KNN), and Bayes were implemented for predicting the product demand. The result indicated SVM as the best-emerging model with an accuracy of 84.8% followed by KNN, and Bayes with 83.6% and 74%, respectively. Another study in [3] demonstrated how transactional data can be used with ML to forecast demand in products in the retail industry. To train the ML models, a sample of 5,115,472 records of receipt data was obtained from the French branch of Belgian supermarket chains' data warehouse. The results revealed that the ML models manage to learn the seasonality effects and allow them to make better predictions. Furthermore, the authors of [4], proposed the use of ML for predicting the agricultural needs of retailers and consumers based on historical data from various warehouses. Three algorithms were implemented, namely, SARIMA, long short-term memory, and Holt-Winters to make the prediction for various SKUs. Based on the findings, different algorithms performed better for different SKUs, however, as a single choice Holt-Winters performed the best on average with the least Mean Absolute Error (MAE) for most of the SKUs. It should be noted that while there are several other metrics used for accuracy evaluation such as mean squared error (MSE), root mean squared error (EMSE), R-squared (R^2), etc, the MAE remains more popular due to several reasons including intuitive interpretation, ease of computation, greater focus on accuracy, etc; therefore, it will be adopted in this study.

Vollmer *et al.* in [5] addressed the challenge of meeting the increasing demand for emergency department (ED) services by implementing ML models to predict the daily number of ED arrivals using additional variables such as seasonal variations, daily weather data and specific planned events. The predictive model, which employed both traditional time series and ML algorithms, demonstrated promising results with a MAE of ±10 to ±14 patients. Moreover, in a study conducted by Tanizakia *et al.* in [6], ML was applied to predict the number of customers at a restaurant on a given day, thereby enabling the restaurant to have its operations and stock adequately prepared to meet the customer demand more effectively. The authors of [7] presented a meta-learning framework based on deep convolutional neural networks for predicting the product sales at the store level. In addition to the historical sales data, they incorporated influential factors such as price, promotions, seasonality and calendar events to make the prediction.

The results indicated that the meta learner that learns features from both the time series of sales and influential factors can improve forecasting performance potentially over the- meta learner using sales time series as the input only.

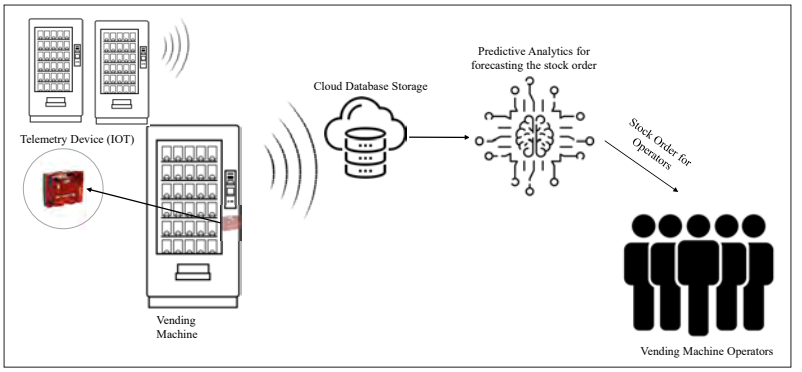


FIGURE 1. The proposed system architecture.

Several studies recommended better performance of the XGBoost model due to its scalability and faster processing speed as compared to the other ML algorithms. Krishna, Akshay *et al.* in their research work [8], compared different ML techniques for retail sales prediction and observed that boosting algorithms outperformed traditional regression techniques. In [9], the authors proposed a sales forecasting strategy based on the LightGBM, SVM and linear regression model. The results of the three prediction models were analysed and it was discovered that the LightGBM model outperformed traditional linear regression and SVM in terms of accuracy. Last but not least, the authors of [10] focused on sales forecasting using three distinct ML algorithms: Gradient Boosting, Random Forest, and K-Nearest-Neighbour.

The Proposed Solution

Unlike all existing studies discussed above, we propose in this study the implementation of ML models to predict vending machine warehouse inventory to cater to the vending machine restocking needs. Figure 1 represents a top-level system architecture of the proposed solution in this work.

More specifically, this figure shows the three main segments of the proposed solution, i.e.,

- Data collection and storage
- Model development
- Model deployment

For data collection, we utilise the IoT telemetry unit that has been installed in the vending machines to capture the transactional data. The data is stored on the cloud database where then the predictive modelling techniques will be implemented to predict the stock order for the warehouse operators.

In this research, we will be using the four different time series forecasting algorithms: XGBoost, ARIMA, Fb Prophet and SVR. The experiment is split into two parts. In the first experiment, we consider the conventional historical data without the external factors to make the prediction whereas in the second part, we add additional variables, first individually and then in combination to assess the impact of adding external factors to the accuracy of the implemented models. The evaluation of the model will be based on the MAE metric.

To authenticate and rigorously test the validity of our proposed methodology, we established a collaborative partnership with a distinguished vending machine enterprise headquartered in the United Kingdom called Broderick's Ltd.

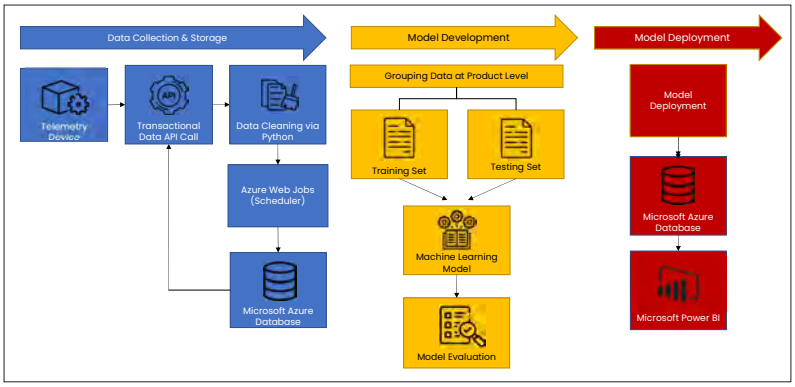


FIGURE 2. The three segments of the considered system: i) data collection and storage, ii) model development and iii) model deployment.

Current Process of Inventory Forecast

Currently, vending machine operators follow set routes to restock machines on a weekly schedule, ordering stock from the warehouse based on their observations and experience. The warehouse initiates stock orders by monitoring depletion using an eight-week rolling average method to forecast upcoming needs. However, operator estimates can be inaccurate, impacting the reliability of stock forecasting. This approach lacks consideration for broader market trends, seasonal variations, and scalability as the business expands.

In this section, we explain the system architecture that we adopted to collect and process the data. The architecture is divided into three segments as can be seen in Fig. 2.

Data Collection and Storage

When the customer makes a product selection on the vending machine, the product details such as product name, price, and sales amount are stored on the telemetry device in the vending machine. This device is the connectivity solution to add real-time data and remote commands to the vending machine. The engine then requests payment authorization and activates the PMT/POS to receive the payment. Once the payment is successfully processed, the telemetry device via the cellular system sends the data to the cloud. This data is fetched through the transactional API which contains the machine data such as vending machine ID, customer name, transaction date, product code, product name, sales amount and sales status.

The transactional API is called to fetch the data, and data cleaning is performed to make the data consistent, accurate, and reliable. The used Python script is scheduled to run on the Microsoft Azure Portal using the Azure Web Jobs to automatically store the data in the Azure database every 10 minutes. An Azure Web Job is a feature provided by Microsoft Azure, a cloud computing platform. It allows running scheduled jobs as part of a web application hosted on Azure App Service. In each run, the transactional data is fetched from the API and is stored on the Microsoft Azure database.

Model Development

The data stored in the database is on-site and machine level. For implementing inventory forecasting at the warehouse level, the data is grouped on the product level to understand the total quantity sold of individual products at a daily level. In this way, all the product data gathered from vending machines is stored in one place and the ML algorithms are implemented centrally regardless of the geographical location of the machine.

Once the data is grouped, it is sorted in ascending order by date for each product since the time series forecasting requires the data to be sequentially ordered. The data is then divided into training and testing sets with a split of 70% and 30% respectively at a product level. The training set is the dataset through which the model learns patterns and features to make the predictions. The testing set is the dataset that is not used during the training phase and is used to evaluate the performance of the algorithms [11].

Once the data is split into training and testing sets at the product level, different machine-learning algorithms are implemented to forecast the product inventory. Once the models are implemented, the model evaluation is conducted to assess the predictive performance of different algorithms and select the model with the best results in terms of the MAE.

After the best model is chosen for making the forecast, the model is deployed onto the Azure cloud and the predictions are stored on the database. The model is scheduled to run weekly to produce a forecast for the next two weeks. The database is connected with Microsoft Power BI, to compare the current stock with the predicted stock and make necessary stock orders for the warehouse.

Data Collection and Preprocessing

In this section, we provide a detailed explanation of our data collection and preprocessing methodology that we adapted to ensure the integrity of the data in order to make accurate predictions. Furthermore, we delve into the ML algorithms that we employed for predicting the warehouse inventory.

Data Source

The transactional data utilised in this study was gathered from the organisation's database hosted on Microsoft Azure. The data was collected for a period from 1st January 2023 to 12th September 2023 for 1,500 snacks and cold drinks vending machines located across all over the UK.

Data Pre-Processing

In order to maintain data consistency and reliability, we conducted data grouping and cleaning on the dataset.

Products Grouping: Initially, the dataset consisted of sales data for individual vending machines. However, as our predictions were intended for the warehouse level, we aggregated the product quantities based on their names, excluding the data related to individual sites, through the use of an SQL query.

Handling Product Data Interruptions: The planograms for vending machines undergo periodic changes, which can lead to interruptions in the continuous recording of product transactions. To ensure an adequate amount of data for all products, we excluded any product that was discontinued during the observation period due to limited quantity information. This exclusion criterion was applied to products that appeared less than 80% of the time within the 255-day period.

Data Features

The data initially provided attributes such as date, product name, and quantity. However, to enhance the prediction accuracy, supplementary explanatory variables were introduced into the model.

1. *Weekday*: We investigated the sales patterns of products across various days of the week to observe any variations associated with weekdays. It was found that the sales vary at different days of the week. To incorporate this specific temporal factor into our dataset, we introduced a new variable called "weekday."
2. *Sales Quantity Deviation Flag*: From time to time, vending machines experience malfunctions and require going through the maintenance procedures. During these maintenance intervals, the machines experience downtime and there is a notable decline in sales. To identify such days, we employ a flag system that assesses whether the average sales on a specific weekday have decreased by 5%. If there is a 5% decrease in sales, the fault flag is set to "1"; otherwise, it is set to "0."
3. *Public Holiday Indicator*: Recognizing that vending machine sales can be influenced by public holidays, we introduced a holiday flag into the dataset. This flag takes on the value "1" to signify a public holiday and "0" when there is no public holiday.

Training and Testing Data

In time series forecasting, it is essential to split the data into training and testing sets for the purpose of evaluating the model performance and fine-tune the models effectively. This division helps in assessing how well the models generalise to unseen future data, preventing overfitting and ensuring realistic performance assessment. Therefore, we split our dataset, dedicating 70% to training and reserving 30% for testing, spanning from January 1st, 2023, to September 12th, 2023. This split ratio was chosen based on considerations such as dataset size, ensuring a balance between training and testing data volumes.

Furthermore, it is worth noting that prior to this partitioning, we upheld the chronological order of the time series data which is a crucial prerequisite for capturing temporal dependencies and facilitating accurate predictions. It should also be highlighted that there is expected variations in the seasonality and trending throughout this period and such variations may affect the performance of the applied models during the testing phase. However, as more data is collected in the future, such variations will diminish.

ML Models

In this study, we employed four different time-series prediction models, i.e., FB Prophet, XGBoost, ARIMA and SVR. Our selection of machine learning models was guided by a combination of factors, including applicability to the problem domain, model complexity, and ease of interpretability, all of which are briefly introduced below.

Facebook Prophet: The FB Prophet is an open-source time series algorithm developed by Facebook's core data science team. The algorithm utilises an additive model that captures non-linear trends, incorporating yearly, weekly, and daily seasonality along with holiday effects [12]. The model allows the inclusion of external factors into the model for making predictions. Moreover, it performs automatic data pre-processing in terms of missing data and handling outliers or capturing fluctuations in the time series.

Extreme Gradient Boosting (XG Boost): XGBoost, a highly popular ML algorithm, is founded upon the principles of the Gradient Boosting Decision Tree (GBDT) technique [13]. In the XGBoost algorithm, multiple decision trees collaborate to make predictions on the dataset. However, what sets it apart is its iterative learning approach. In each iteration, XGBoost focuses on rectifying the mistakes or residuals made by the previous model. It pays special attention to instances that were poorly predicted, thus continuously enhancing its predictive capabilities. To ensure the model doesn't become excessively complex and prone to overfitting, XGBoost incorporates regularisation terms into the loss function during the training process. This regularisation effectively controls the model's complexity, resulting in a more generalised and robust predictive model.

Auto-Regressive Integrated Moving Average with Exogenous Variables (ARIMAX): ARIMA/ARIMAX is a widely employed statistical model utilised for forecasting time series data. ARIMA decomposes the time series analysis into three essential components. First, the autoregressive component (AR), denoted as (p), assumes that the present value of the time series relies on its past values. Second, the integrated or differencing order (I), denoted as (d), transforms the time series data into a stationary form, enabling the model to capture changes in the series values over time. Lastly, the moving average (MA) component, represented as (q), deals with forecasting errors by considering that the current value of the time series is influenced by the residual errors from prior observations. This technique uses the time series principles to predict future outcomes as a linear equation based on input data and taking prediction error into account [14]. The model also considers external factors that are called exogenous variables to help analyse and identify the hidden patterns in the data for making accurate predictions.

SVM Regression Rephrase: The SVM, developed by Vapnik and others in 1995, is used for many ML tasks such as pattern recognition, object classification, and in the case of time series prediction, regression analysis [15]. Unlike conventional regression techniques, which strive to minimise the discrepancy between predicted and actual values, SVR directs its efforts towards discovering a regression line, often referred to as a hyperplane, that effectively accommodates the training data within a predefined margin or tube. This margin or tube serves as a buffer, signifying the permissible level of divergence from the actual values. In the training phase, SVR aims to identify the hyperplane that best fits the training data while staying within a predefined tolerance margin. Data points that fall outside of this margin are referred to as "support vectors." These support vectors play a crucial role in measuring how far the margin boundaries deviate from them, thus

Algorithm	Parameter Settings	

XG Boost	n_estimators = 100;
	Learning_rate = 0.1;
	sub_sample=0.8;
	reg_lambda=0.1
		colsample_bytree= 0.8
		Max_Depth = 5
		reg_alpha=0.1

FB Prophet	Interval_witdth = 0.95 Seasonality Mode: Multiplicative n_changepoints = 25	
SVR	Kernel: RBF; Gamma: Default (Auto);	C: 1.0 Epsilon: 0.1
ARIMAX	Autoregressive Order (p): 1 Integrated Order (d): 0 Moving Average Order (q): 0 Exogenous Variables: Weekday, Public holiday, Sales deviation flag.	

TABLE 1. Configurations employed for each of algorithm.

	Scenario #1	Scenario #2			
	No external variable	Weekday	Public holiday	Sales deviation flag	All external variables
XG Boost	46.13	39.2	42.15	26.6	22.7
FB Prophet	38.8	38.8	39.5	38.0	37.7
ARIMA(X)	41.60	40.1	41.62	30.26	26.9
SVR	40.02	40.02	42.42	32.1	37.8

TABLE 2. The results obtained from scenarios #1 and #2. All values represent the MAE.

impacting the calculation of errors associated with the margin. SVR has several advantages, including the ability to handle nonlinear relationships, robustness to outliers, and the ability to control the complexity of the model through the margin parameter. However, it should be highlighted that SVR can be computationally expensive, especially for large datasets, and requires careful tuning of its parameters for optimal performance. The configurations employed for each of the algorithms are illustrated in Table 1.

Result and Analysis

In this section, we present, discuss and compare various results. Two scenarios are considered herein to make the inventory predictions for the warehouse. Scenario #1 comprises of implementing the four ML algorithms without integrating external variables whereas scenario #2 consists of adding external variables, first individually and, subsequently combined. The external variables include weekdays, public holidays, and sales deviation flag.

To evaluate the performance of the models, the MAE for each model was considered. MAE is a statistical metric used in ML to measure the average absolute difference between predicted and actual values. The formula for calculating MAE is as follows

$$\text{MAE} = \sum_{i=1}^{n} |y_i - x_i| \quad (1)$$

where n is the number of data points, y_i is the predicted value, x_i is the actual value, and Σ and $|.|$ represent the summation and the absolute value operators. The evaluation metric MAE is chosen due to several reasons including intuitive interpretation, ease of computation, greater focus on accuracy, and resilience against the outliers in the dataset.

Based on the results as in Table 2, in the first scenario (no external variable), the FB Prophet algorithm had a superior performance as compared to all the other algorithms with an MAE of 39 units/day. This means that on average the predicted quantity of a product will deviate by 39 units on average. The second best-performing model was SVR followed by ARIMA and XGBoost with the MAE of 40, 41.6 and 46.13, respectively.

In Fig. 3, we present the actual and predicted quantities excluding additional variables for different time series models. It is interesting to see from this figure that the FB prophet's prediction outperforms the other models. This can be justified by the fact that FB Prophet utilizes a robust modeling framework that can handle outliers and missing data more effectively compared to some traditional ML and statistical models such as XGBoost, SVR, and ARIMA. This robustness helps improve the stability and reliability of predictions, particularly in noisy or uncertain environments. However, the difference between the actual and predicted quantities establishes that the predictions are not promising and cannot be utilised in a practical scenario. Additionally, it is apparent that the ARIMA model fails to adequately capture the inherent patterns and fluctuations within the time series data, resulting in a noticeable constant line.

In scenario #2, we systematically introduced external variables into our ML algorithms, first individually and subsequently in combination to assess the impact on the performance of the models. Table 2 presents the experiment results.

The first external variable that was added to the model is the day of the week, a significant factor due to the observed daily sales fluctuations across different days of the week. Based on the results, all the ML models experienced a slight improvement in their MAE. The second variable that was introduced is the public holiday. Given the site dynamics, some locations such as offices and educational institutions experience reduced or zero sales due to public holidays, while others such as shopping and leisure centres witness an increase in sales. Using public holidays as the only external variable, it is clear that the performance of the models improves slightly except for SVR, which exhibited a slight decline in the MAE. The third external variable that we included in the model is the sales deviation flag. Due to machine faults, the sales of the machines can significantly deviate, impacting the overall quantity of products sold. Integrating this flag into our ML models yielded improvements across all algorithms. Notably, XGBoost showcases the most significant improvement bringing the MAE down from 46 to 26 units/day and for ARIMAX bringing the MAE down from 42 units/day to 30 units/day. This is perhaps because XGBoost employs an ensemble learning technique known as gradient boosting, which combines multiple weak learners (decision trees) to create a strong predictive model.

Overall, the inclusion of external variables individually, one at a time led to noticeable enhancements in model performance, emphasising their significance in mitigating predictive errors. This step is important as it can assess the relevance and impact of each variable on the prediction task. Adding the external variables individually

provided supplementary context and explanatory power that helps the model better discriminate between different classes or categories in the dataset. This additional information can lead to more accurate predictions by enabling the model to better capture the underlying patterns and relationships in the data.

We then included all the external variables i.e., weekdays, public holidays, and sales deviation flags into our ML models. Based on the results XGBoost emerged as the top-performing model with an MAE of 22 units/day, followed by ARIMAX, FB Prophet, and SVR. Figure 4 presents the actual and predicted quantities including additional variables for different time series models. From the graph, we can see that the performance of the XGBoost has significantly increased as can be seen by the difference between the actual and predicted quantities. Also, it can be observed that as the time progresses, the predictions are improving over time. In contrast, the performance of FB Prophet and SVR does not exhibit substantial improvement. Also, it is apparent that the best performance for SVR is achieved when incorporating the sales deviation flag only. This emphasised on the significance of customising external variables to align with the unique requirements of the model, thereby enhancing predictive accuracy.

It is worthwhile highlighting that decentralised or federal learning (FL) can offer some advantages over the considered centralised approach. For instance, FL can ensure that customer purchase data, which might contain sensitive information, remains on the local machine, which reduces the risk of data breaches and complies with privacy regulations such as GDPR. Also, with FL only model updates will be transmitted, not raw data, significantly reducing the amount of data sent over the network. This is particularly beneficial for vending machines with limited or expensive connectivity. Furthermore, vending machines in different locations can learn demand patterns specific to their contexts, enhancing prediction accuracy. In terms of operational resilience, the failure of a single vending machine will not compromise the overall learning process, enhancing resilience; machines can also operate and update models independently without needing constant connectivity to a central server.

Challenges and Future Directions

The integration of IoT and ML in vending machines involves the collection and analysis of vast amounts of data. Ensuring the security and privacy of this data poses a significant challenge, as unauthorised access could compromise sensitive information about customer preferences and transactions. Furthermore, integrating IoT and ML technologies into existing vending machines can be costly, especially for small and medium-sized businesses. The initial investment in upgrading machines and implementing the necessary infrastructure may be a barrier for some operators. Future research directions on this topic can include:

- Enhanced predictive analytics: Future research can focus on refining ML algorithms for more accurate predictive analytics. This includes developing models that can better anticipate customer preferences, optimise inventory management, and adapt to dynamic market conditions.

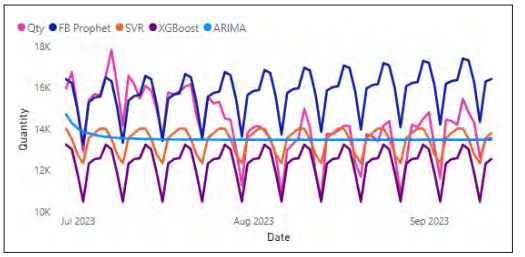


FIGURE 3. CWarehouse inventory prediction without the external variables.

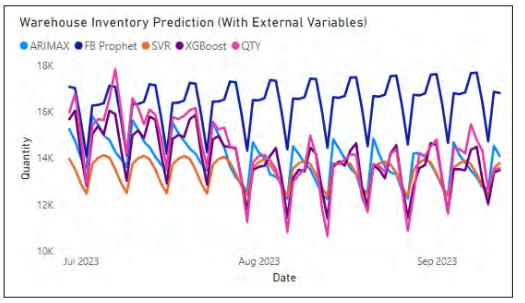


FIGURE 4. The actual and predicted quantities for each of the models using all the external variables.

- User experience and human interaction: Investigating ways to improve user experience through human-machine interaction is crucial. This includes studying user acceptance, designing intuitive interfaces, and understanding how customers interact with smart vending machines to enhance overall satisfaction.
- Blockchain for security: Research into integrating blockchain technology for secure and transparent transactions in vending machines can address data security concerns.
- Decentralised learning algorithms: As discussed in the previous section, FL can offer interesting insights into the system performance at machine level as compared to the decentralised approaches adopted in the current study.

Conclusion

This article proposed an IoT-driven ML-aided solution to predict inventory for a vending machine warehouse to enable the timely ordering of products with the goal of effectively replenishing the vending machines, achieving cost effectiveness, waste reduction, resource allocation optimization and ensuring adequate fulfilment of consumer demand. Four different time series forecasting algorithms were adopted, namely, XGBoost, Fb Prophet, ARIMA and SVR. The results showed that XGBoost is the best performing model with the lowest MAE of 22.7 followed by ARIMAX, FB Prophet and SVR. Furthermore, our investigation indicated the pivotal role of the external variables in enhancing the performance of the considered ML models. Also, several current challenges and future research directions have been discussed. In future research, we implement more advanced time series models such as SARIMAX which have potential for enhanced predictive power and more advanced capabilities in capturing time-varying relationships using the exogenous variables compared to the ARIMA model used in this study.

References

[1] R. Sengar and Dr. Ahmed, "Review on Trends in Machine Learning Applied to Demand & Sales Forecasting," *Smart Moves Journal IJOSCIENCE*, vol. 5, pp. 4, Dec. 2019.

[2] A. Mishra and M. Mohapatro, "Real-Time RFID-Based Item Tracking Using IoT & Efficient Inventory Management Using Machine Learning," *2020 IEEE 4th Conf. Information & Communication Tech. (CICT)*, Chennai, India, 2020, pp. 1–6.

[3] K. B. Agbemadon, R. Couturier, and D. Laiymani, "Overstock Prediction Using Machine Learning in Retail Industry," *2023 3rd International Conference on Computer, Control and Robotics (ICCCR)*, Shanghai, China, 2023, pp. 439–44.

[4] N. P. Kumar *et al.*, "Machine Learning Based Predictive Analytics For Agriculture Inventory Management System," *2022 4th Int'l. Conf. Cognitive Computing and Information Processing (CCIP)*, Bengaluru, India, 2022, pp. 1–7.

[5] M. A. C. Vollmer *et al.*, "A Unified Machine Learning Approach to Time Series Forecasting Applied to Demand at Emergency Departments," *BMC Emergency Medicine*, vol. 21, 2021, p. 9.

[6] T. Tanizaki *et al.*, "Demand Forecasting in Restaurants Using Machine Learning and Statistical Analysis," *Proc. 12th CIRP Conf. Intelligent Computation in Manufacturing Engineering*, Gulf of Naples, Italy, July 2018.

[7] S. Ma and R. Fildes, "Retail Sales Forecasting with Meta-Learning," *European Journal of Operational Research*, vol. 288, no. 1, 2021, pp. 111–28.

[8] A. Krishna *et al.* "Sales-Forecasting of Retail Stores Using Machine Learning Techniques," *3rd IEEE Int'l. Conf. Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, 2018.

[9] Qiao, Zihan. "Walmart Sale Forecasting Model Based On LightGBM," *2nd IEEE Int'l. Conf. Machine Learning, Big Data and Business Intelligence (MLBDBI)*, 2020.

[10] G. Chandrashekar and F. Sahin, "A Survey on Feature Selection Methods," *Computers & Electrical Engineering*, vol. 40, no. 1, 2014, pp. 16–28.

[11] G. S. Handelman *et al.*, "Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods," *AJR Am J Roentgenol*, vol. 212, no. 1, Jan. 2019, pp. 38–43.

[12] S. J. Taylor and B. Letham, "Forecasting at Scale," *PeerJ Preprints*, 2017, pp. 5:e3190v2.

[13] T. Chen and C. Guestrin, "Xgboost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD Int'l. Conf. Knowledge Discovery and Data*.

[14] C. Deb *et al.*, "A Review on Time Series Forecasting Techniques for Building Energy Consumption," *Renewable and Sustainable Energy Reviews*, vol. 74, 2017, pp. 902–24.

[15] V. N. Vapnik, 1995, The Nature of Statistical Learning Theory. Springer, New York; http://dx.doi.org/10.1007/978-1-4757-2440-0.

Biographies

Umair Mehmood is the founder and CEO of Datapre8, a UK-based data analytics firm. He holds a Master's degree in Data Science from the University of South Australia, where he received the prestigious Vice Chancellor's Scholarship. Umair was awarded distinction for his contributions to the Knowledge Transfer Partnership (KTP) project, a collaboration between Manchester Metropolitan University and Broderick Group. Additionally, he has authored a research paper and is recognized for his expertise in machine learning and data-driven strategies.

John Broderick's biography was not available at the time this issue went to press.

Simon Davies's biography was not available at the time this issue went to press.

Ali Kashif Bashir is a Professor at the Department of Computing and Mathematics, Manchester Metropolitan University, UK. He is the director of Secure and Intelligent Systems Research Group; Future Networks Lab and IoT/Cybersecurity Testbed. He is a Fellow of the Royal Society of Arts, a senior member of IEEE, a member of several technical societies, and received the Clarivate Highly Cited Researcher Award in 2023.

Khaled M. Rabie (K.Rabie@ieee.org) received the M.Sc. and Ph.D. degrees in electrical and electronic engineering from the University of Manchester, in 2011 and 2015, respectively. He is currently a Reader with the Department of Engineering, Manchester Metropolitan University (MMU), UK. His current research interests focus on designing and developing next-generation wireless communication systems. He is a Fellow of the Higher Education Academy (FHEA) and a Fellow of the European Alliance for Innovation (EAI).

IEEE GLOBECOM®

IEEE ComSoc®

IEEE Global Communications Conference 2024

8–12 December 2024 // Cape Town, South Africa

*Be a part of the premier conference on communications technology and innovation! Register by **6 November 2024** to take advantage of early bird savings!*

PROGRAM HIGHLIGHTS:

- Keynotes
- Technical Symposia
- Industry Panels & Presentations
- Workshops
- Tutorials
- Women in Communications Engineering Program
- Young Professionals Program
- Exhibitions and Demos
- Humanitarian Technologies Board Events

For more information, visit **ieee-globecom.org**

Efficient Transformer-Based Hyper-Parameter Optimization for Resource-Constrained IoT Environments

Ibrahim Shaer, Soodeh Nikan, and Abdallah Shami

Abstract

The hyper-parameter optimization (HPO) process is imperative for finding the best-performing Convolutional Neural Networks (CNNs). The automation process of HPO is characterized by its sizable computational footprint and its lack of transparency; both important factors in a resource-constrained Internet of Things (IoT) environment. In this article, we address these problems by proposing a novel approach that combines transformer architecture and actor-critic Reinforcement Learning (RL) model, TRL-HPO, equipped with multi-headed attention that enables parallelization and progressive generation of layers. These assumptions are founded empirically by evaluating TRL-HPO on the MNIST dataset and comparing it with state-of-the-art approaches that build CNN models from scratch. The results show that TRL-HPO outperforms the classification results of these approaches by 6.8% within the same time frame, demonstrating the efficiency of TRL-HPO for the HPO process. The analysis of the results identifies the main culprit for performance degradation attributed to stacking fully connected layers. This article identifies new avenues for improving RL-based HPO processes in resource-constrained environments.

Introduction

Convolutional Neural Networks (CNNs) are the staple implementation of neural networks (NNs) for image classification and object detection. The progress in this field is attributed to the increased complexity of CNN architectures in terms of the type of connections among different layers and their depth, which increased the computational demands [1]. Therefore, the accuracy of these models depends on the number and the type of layers, the connections between these layers, and the parameters assigned to each layer [2]. Other important factors dictating the design of CNNs include training time, inference time, and energy consumption, emphasized in the resource-constrained Internet of Things (IoT) environment. The number of possible choices for designing CNNs is extremely large, which promoted the automated CNN architecture search, achieved through Neural Architecture Search (NAS).

The NAS field has recently seen great progress, due to the incorporation of Reinforcement Learning (RL) agents to search for the best CNN configurations. The RL's appeal stems from its generalizability of hyper-parameter (HP) combinations via function approximation and the trial-and-error approach, which can reduce the computational demand of NAS [3]. This field is dominated by the literature that either builds CNNs from scratch using basic NN layers or a pre-defined collection of these layers or optimizes the HPs of an already-existing CNN.

Autonomous vehicles (AV) are part of the envisioned IoT applications that utilize edge servers of limited computational resources. The realization of AVs heavily depends on image classification models deployed on these servers. Their disparate computational capabilities are prohibitive for the prolonged execution of large models. Model partitioning is key to resolving these limitations toward fulfilling the promise of AVs [4]. This requires gathering insights into the contribution of each CNN model's layers to the classification results. In current implementations, this vision is stalled by multiple oversights. The sequential nature of model generation results from the dependence between layer combinations, which can be reflected by the recurrent structures of Recurrent Neural Networks (RNNs). Therefore, RNN-based controllers constructing CNN models dominated the RL-based NAS field [5, 6]. The one-at-a-time processing of inputs inhibits their parallelization, resulting in their insurmountable computational footprint that is ill-fitted for the resource-scarce vehicular environment. On a different note, these controllers are built to receive a reward when a terminal condition is encountered, masking the contribution of each layer to prediction results. Within the constraints of computational resources and model partitioning requirements, the state-of-the-art (SOTA) approaches are limited by their impracticability and lack of transparency, hindering their deployment in real-world environments.

This article presents a Transformer-based Reinforcement Learning Hyper-parameter Optimization (TRL-HPO) model that alleviates the shortcomings of the current methods. The TRL-HPO addresses the transparency and long-computational times of RL-based solutions with competitive performances. The Transformer architecture overshadows the RNN-based meth-

Digital Object Identifier: 10.1109/IOTM.001.2300285

The authors are with the University of Western Ontario, Canada.

ods by integrating the multi-headed self-attention (MHSA) mechanism that enables parallelization [7], addressing the long computational times. The TRL-HPO controller equipped with the attention mechanism allows us to gain insights from the layer-generation process, enhancing its interpretability. The reward is produced with every generated layer to showcase improvement instead of waiting for termination conditions to unfold, a condition matching CNN partitioning requirements, as depicted in Fig. 1. The contributions of this article are as follows:

- Propose a novel Hyper-parameter Optimization (HPO) process named TRL-HPO that is the first to combine transformer architecture and actor-critic (AC) RL;
- Enhance the CNN's model performance generated within a shorter period compared to SOTA approaches using the TRL-HPO process;
- Improve the transparency of the CNN model generation process by examining the attention-reward affinities and their layer combinations;
- Create open challenges related to RL-based HPO processes, including TRL-HPO, for researchers to address.

The rest of the article is organized as follows. We present the related work and its limitations. We detail the proposed framework. We explain the implementation details and evaluation criteria. We analyze the obtained results. We conclude the article.

Related Work

The field of hyper-parameter optimization (HPO) is an important research topic in Machine Learning (ML) with practical implementations in real-world environments that enhance the performance of Deep Neural Networks (DNNs).

Bayesian Optimization (BO), Evolutionary Search (ES) algorithms, and RL agents are the three main tools for HPO implementation. BO and ES methods are limited by their assumptions [2] and lack of generality [8], which favors RL techniques. The works of Baker *et al.* [6] and Zoph *et al.* [5] are the first to propose incorporating RL methods into generating CNN models. The former work utilized RNN-based controllers using an off-policy RL algorithm to sample CNN architectures while the latter utilized a value-based Q-learning approach. The results of these seminal works highlighted the trade-off between the running time of HPO methods and the accuracy of the obtained models.

The computational footprint of RL implementations for HPO promoted the utilization of multi-agent RL that shrinks the state space of each agent. The work of Neary [9] uses multiple agents to optimize HPs of the CNNs built from scratch, whereby a master agent orchestrates the synchronization between the other agents outputting HPs. On the other hand, the work in [10] assigns each agent to optimize the HPs of an already-existing CNN layer, such that the dependence in HP space is mapped using a shared Q-table between consecutive layers. While all of these studies focus on a single objective, the works presented by Hsu *et al.* [11] and Tan *et al.* [12] incorporate multiple objectives in the reward function formulation. MONAS [11] considers the energy and accuracy constraints, whereas MNASNet [12] integrates the inference latency of developed models.

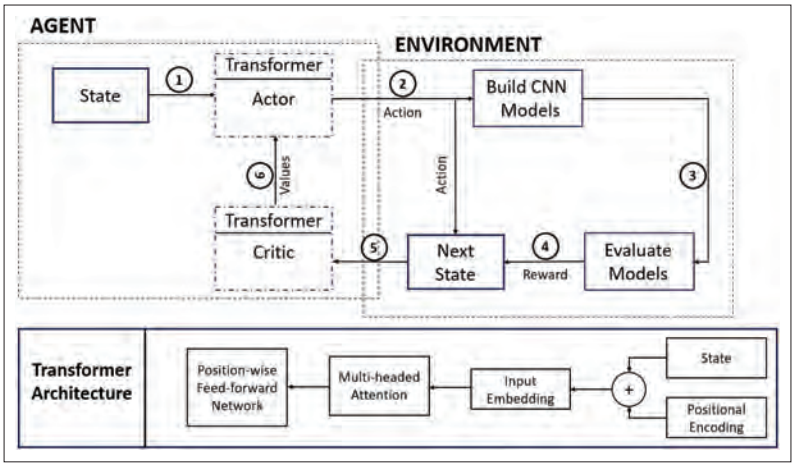


FIGURE 1. TRL-HPO framework.

The shortcomings of RL integration into the HPO process hinder their deployment in IoT environments. First, the adopted models lack the transparency that shows the interdependence between different layers. Second, the long convergence times are prohibitive for deploying these models in resource-constrained environments. Third, some works focus on optimizing the HPs of a specific layer, reducing the state space at the expense of the layer's diversity. Lastly, many works include prior knowledge in the layer-generation process, such as the addition of dropout layers [5, 6]. This article addresses these limitations by proposing a transformer-based RL controller and the reward function formulation. The MHSA facilitates the training process and adds transparency to the model generation by analyzing the attention values. On the other hand, the reward function reflects the contribution of each layer to enhancing the classification results, which favors CNN model partitioning.

Methodology: TRL-HPO

This section explains the building blocks of TRL-HPO, including the transformer and RL's AC architectures, which are shown in Fig. 1, and the motivation for this combination toward a more efficient and transparent HPO process.

Transformer

The vanilla transformer is a sequence-to-sequence model, which avoids the recurrence structures of RNNs by integrating the innovative self-attention mechanism [7]. This architecture enables parallelization accelerating the training of transformer models. The building blocks of transformers include attention mechanism, multi-headed attention (MHSA), position-wise feed-forward network, and positional encoding (PE). The transformers follow the encoder-decoder structure thereby each of these structures is formed by stacking combinations of these identical layers, as depicted in Fig. 1.

The PE computes the sequence's order and is added to the input embedding of the encoder and decoder stacks so that the order is incorporated into the input and output data. To obtain a unique order, a sinusoidal function inputs the position and the embedding dimensions. The self-at-

The generated CNN models are constrained to 6 layers, formed as a combination of the three basic layers: CNN layer, FCL layer, and Maximum Pooling (MaxPool) layer.

tention module calculates the weights each input sequence assigns to other sequences. This way the self-attention reflects the dependencies between input and specific output and improves the modelling of long-range dependencies [7]. As such, the attention mechanism resembles a fully connected layer (FCL), whereby the weights reflect the pairwise relationship from previous inputs. To fully exploit this method, the transformer includes MHSA that splits the attention calculation among different heads of the embedding that can be calculated independently, facilitating its parallelization. Lastly, a feed-forward neural network (FFNN) is the subsequent layer to the MHSA.

Actor-Critic Reinforcement Learning

RL teaches an agent to perform a task by accumulating experiences from interacting with its environment. Its actions are refined based on a reward function that signals the utility of these actions [13]. The RL's experimentation with the absence of prior knowledge mirrors the ignorance of the best combination of basic CNN layers. The only knowledge for the HPO process relates to the order of the stacked layers that starts with grid-like inputs and ends with an FCL. The combination of unknown orientation and trial-and-error experimentation matches the requirements of the HPO process.

The RL methods can be split into value-based and policy-based methods. The value-based method estimates the quality of a state-action pair using a value function and optimizes this value iteratively using actions that maximize it. The high variability of value estimations and their under-performance in scenarios with continuous action spaces promotes policy-based methods [13]. The policy-based methods optimize the policy to maximize the expected cumulative rewards [13]. However, the latter methods suffer from sample inefficiency, which is an asset of value-based methods. To complement these methods' advantages, an actor-critic (AC) approach combines both policy-based (actor) and value-based (critic) methods [13]. Projected to the HPO problem, the actor outputs an action that maps to a layer and its HPs. The critic evaluates this action by outputting a value showing the action's quality [13].

Transformers and Actor-Critic RL

The proposed framework, Transformer-based Reinforcement Learning HPO (TRL-HPO) is the convergence of transformers and AC RL. The actor and the critic are implemented using the transformer's decoder architecture. The steps involved in the actions' generation, CNN model construction, and their evaluation are depicted in Fig. 1. The combination of TRL-HPO enables the RL agent to harness the strengths of transformers to benefit the HPO process. This new framework is the **first trial** to integrate transformers into the HPO process, opening a new frontier toward the exploitation of this novel architecture to address a lingering problem in the field of ML. This experiment evaluates the viability and suitability of this architecture to the HPO process.

We analyze TRL-HPO based on its inherent advantages and benefits versus the SOTA approaches. Two main advantages are reaped. The first is that TRL-HPO builds models from scratch, rendering it a general-purpose method. The proof of concept is constrained to a single use case; however, the framework can be expanded to any DNN-related problem. The second concerns the definition of the reward function that is progressively updated with each stacked layer. While this definition incurs an extra computational footprint by training progressively deeper models, it facilitates understanding the contribution of each layer. This feature fits the requirements of environments that seek to balance resource constraints and accuracy objectives.

The transformer architecture results in two main advantages to the HPO process. The first relates to handling long sequences, which benefits the transferability of TRL-HPO from small to larger datasets. The second benefit relates to the MHSA, which facilitates the parallelization of TRL-HPO and reduces its running time. Additionally, MHSA reveals the relationships between generated layers, which allows the inference of the combination of layers that produce better results, enhancing the transparency of the HPO process.

Implementation Details

The evaluation of TRL-HPO is conducted on the MNIST dataset [14], which is a large database of handwritten digits, from 0 to 9, containing 60,000 training and 10,000 testing images. Each image is a greyscale image of 28 × 28 dimensions. The availability and limited complexity of MNIST can serve as a good proof-of-concept for TRL-HPO. The former factor enables a fair comparison with other SOTA approaches, such as [6, 9]. The latter factor facilitates training CNN models in a limited time; an advantage when working with hardware constraints. The comparison with SOTA approaches is based on the convergence time to the best solution (hrs), the classification accuracy (%), and the best CNN classification model (%) obtained by each SOTA approach upon the completion of TRL-HPO process (**AccTime**).

Reproducibility is a major issue that plagues the HPO process [2]. With this factor in mind, the choice of competing methods depended on two conditions:

1. The availability of source code that enables a fair comparison with TRL-HPO and dispels any introduced biases by implementing a method from scratch.
2. The similarity in the experimental procedure for building CNNs from scratch using an RL-based method and the availability of results applied to the MNIST dataset. Therefore, we compare TRL-HPO with Baker *et al.* [6] and Neary [9], which abide by at least one of these conditions.

The generated CNN models are constrained to 6 layers, formed as a combination of the three basic layers: CNN layer, FCL layer, and Maximum Pooling (MaxPool) layer. The set of HPs used in the experimental procedure is summarized in Table 1. Stacking a combination of these layers adapts to the input structure. The state and action space design and the reward function are imperative to drive the RL agents' action generation, which requires defining the transformers' input sequence.

The reward function represents a layer's contribution towards improving the classification results, which means that the DNN model's first layer produces the highest rewards compared to any subsequent layers. Each layer is represent-

ed as a combination of the representation of the layer itself and its HP, and the performance of the obtained model on the MNIST's validation set. The generated layers' representation is obtained from the output layer of the RL's actor, summarized in four values. These values represent the action space mapped first to a layer and then to that layer's HPs. The model's performance is represented with 32 values, each calculating the model's accuracy results on each validation set's batch of size 16. The heterogeneity in the dimensions of the action space and the performance requires a mapping to a uniform representation. This goal is achieved using a static NN that takes these inputs and produces a uniform output of 64 values, referred to as Intermediate Model Representation (IMR). When a layer and its performance are obtained, the state space, reflecting the current state of the environment, should also change. Since the state space represents a sequence of layers, the index corresponding to the generated layer is updated with the IMR. With more layers, each index is updated with a new representation. This way the state space reflects the two important pieces of information for each layer, its HPs, and its performance. This process was touched upon in Fig. 1.

The stopping criteria for the RL agent involve generating more than 6 layers, minimal improvement in the performance with the addition of layers (0.001), or accuracy below 60%. These criteria are defined to avoid any unnecessary generation of models. A Deep Deterministic Policy Gradient (DDPG) [13] handles the continuous action spaces in this environment. The critic and the actor consist of target and online subnets to avoid radical changes in critic and actor updates [13]. In DDPG, the agent explores by adding random noise to the sampled actions, so that the agent can experiment with different combinations of layers and obtain their performance results. To remove data correlations for the RL agent inherent in the sequential structure of the data, experience replay (ER) [13] buffer is utilized to store data during the exploration stage.

The actor and the critic follow the transformer architecture, depicted in Fig. 1. Two main differences highlight these components' distinctive roles: the transformer's output and the learning rate (lr). The actor's output is in the [0, 1] range mapped to a layer, while the critic's is in the [–1, 1] range representing the state's Q-value. Regarding the lr, it is recommended that the actors' (1e-5) be slower than the critics' (1e-4) [13]. The input embedding's dimension for each input space is equal to 64. The embeddings are inputs to multiple encoding layers, equivalent to 2 in our implementation, each constituting a Transformer block, represented by the MHSA and FFNN. Each of these blocks has the same number of input and output dimensions. The definition of MHSA and FFNN requires highlighting two values, **number of heads** equal to 4 and **expansion factor** equal to 4.

The training process is realized over several episodes. In each episode, 10 full models are generated via CPU parallelization. Once the ER buffer is full, five RL optimization rounds are implemented in each episode. The number of episodes and the size of the ER buffer are important parameters for the HPO process. Their optimization is imperative to obtain models with good performances in a short time. During the HPO process, the training set consists of 20,000 images while the validation set includes 10,000 images out of the original 60,000 images. The experimental procedure is conducted on SHARCNET's Graham cluster on a node with one V100 GPU, 64 Gbs of RAM, and 12 cores, such that each experiment does not exceed 7 days to avoid long queuing times. The implementation is available on the GitHub repository.[1]

Layer	HP	Values
CNN	*filters*	{8, 16, ..., 128}
	kernel	{3, 5, 7}
	stride	{1, 2, 3}
FCL	*neurons*	{16, 24, ..., 512}
	bias	{T, F}
	activations	{None, relu, leakyrelu, tanh, sigmoid, elu, gelu}
MaxPool	*kernel*	{2, 3, ...,, 8}
	stride	{1, 2, 3}
	padding	{0, 1, 2, 3}

TABLE 1. Set of hyper-parameters.

Methods	Best Accuracy (%)	Running Time (hrs)	AccTime (%)
TRL-HPO	99.1	99.3	**94.6% ± 3.8%**
MetaQNN [6]	**99.5**	192–240	88.5% ± 4.3%
Neary [9]	95.8	**1.78**	95.8%

TABLE 2. TRL-HPO versus SOTA in terms of accuracy and running time.

Results and Discussion

This section reports and analyzes the results based on the defined performance metrics and layer-generation interpretability. TRL-HPO results are obtained in the exploitation phase of the offline RL.

TRL-HPO vs. SOTA

Table 2 shows the results of TRL-HPO versus SOTA methods that built CNN models from scratch. Based on evaluation results, **MetaQNN** produces the best model when classifying MNIST data reported in the **exploration phase**. This result is due to the controller's freedom to generate models up to 12 layers and the integration of priors by adding dropout layers [6]. The accumulation of more layers improves classification results; however, at the expense of wider search space that increases the running time. This is evident with the computational footprint of **MetaQNN** that exceeds TRL-HPO, despite running on 8× the amount of GPU resources. On the other hand, the approach in [9] is the first to converge based on the reported results. However, this method experimented with four configurations, highlighting the limited exploration of this strategy and the pre-mature convergence of the master agent. This limitation is reflected by its poor best models compared to the other two approaches, **excluding** it from further analysis. Since the larger search space of **MetaQNN** would not provide a fair comparison with TRL-HPO, it was imperative to compare these two approaches using **AccTime**. Within TRL-HPO convergence time, its generated

[1] https://github.com/Western-OC2-Lab/TRL-HPO.

In a resource-constrained IoT environment, analyzing the energy consumption and processing power of TRL-HPO compared to SOTA methods is imperative.

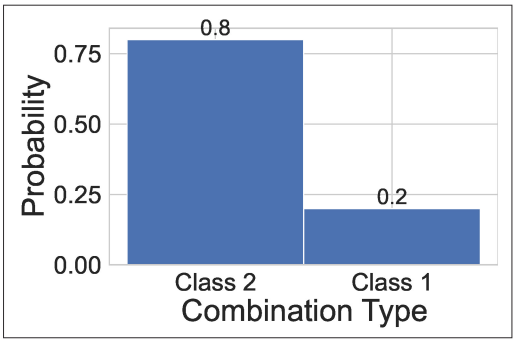


FIGURE 2. Layer affinity for models with negative rewards.

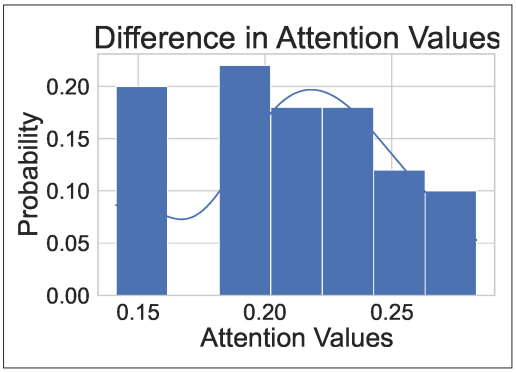


FIGURE 3. Differences in attention between layers of best reward and layers with negative reward.

models exhibit the best performance compared to the SOTA models. Despite TRL-HPO's need to generate more models given the requirement of progressive reward, it outperforms other models in the **AccTime** metric. Compared to other SOTA methods, TRL-HPO can progressively generate CNN models with satisfactory performance in a shorter period. These two conditions are important for resource-constrained IoT environments. In terms of **AccTime**, on average, TRL-HPO outperforms **MetaQNN** by 6.8% in accuracy. The improvement in the accuracy result of TRL-HPO compared to **MetaQNN** is statistically significant (p-value = 0.02) by applying a t-test on the distribution of TRL-HPO versus the 1-sd values of **MetaQNN**.

In a resource-constrained IoT environment, analyzing the energy consumption and processing power of TRL-HPO compared to SOTA methods is imperative. We analyze these methods using the method's number of parameters and the floating point operations (FLOPs), which cover IoT-related concerns. The **MetaQNN** requires storing all models and state transitions to find the best-performing models, which with many combinations of layers and HPs, is prohibitive for the IoT environment. In Neary *et al.* approach [9], an RL agent is assigned to every HPs and a master agent finds the best combination of individual HPs. For Neary *et al.*, 399K FLOPs are required for inference whereas TRL-HPO requires 991k FLOPs. On the other hand, both models have the same number of model parameters (78.8k parameters).

Layer Analysis

Investigating the effect of the addition of layers is central to understanding the contribution of each layer towards improving accuracy results and the combination of layers that yields the best performances. This analysis unveils the complex relationships between layers facilitating the transparency of the HPO process. Both goals were considered in the design of TRL-HPO, representing a key differentiating factor versus SOTA methods. To gain these insights, two questions need to be answered.

1. What are the layer combinations degrading the performance of CNN models?
2. How are layer affinities reflected in the attention mechanism?

Figure 2 summarizes the distribution of layer combinations with negative rewards. Two combinations stand out, **Class 1** represented by Conv2D and Conv2D combinations and **Class 2** represented by two consecutive FCLs. A layer combination refers to the layer that produced a negative reward and its previous layer. Two main insights can be gathered from Fig. 2. The first is that the negative reward is overwhelmingly attributed to the accumulation of FCLs. This means that the stacking of multiple FCL layers is superfluous on the MNIST dataset, a conclusion that aligns with the best m yers can degrade model performance, especially with MNIST data of limited visual complexity. As such, these observations are beneficial to garner knowledge about stacking of layers, suitable for environments with resource constraints.

Figure 3 depicts the distribution of the differences in attention values between layers with the best performances and layers with negative rewards. The distribution demonstrates the close fidelity of the attention mechanism and the utility of a specific layer in its classification performance. As such, the attention mechanism captures the close affinity between layers assigning higher weights to promising layers while shunning layers with poor performances. The difference in attention values averages 0.22, a significant value considering that some attention is assigned to the layers yet to be generated. In these cases, the attention values of the layer degrading model performances are higher than these to-be-generated layers. Therefore, we conclude that the attention mechanism reflects the relationship between layers within the context of generating best-performing models.

Open Challenges and Future Recommendations

The implementation of the TRL-HPO process opens new avenues for exploring the improvement of this process, which spans many research and practical questions about RL-based HPO implementations, summarized in Table 3.

Computational Time

Despite improvements in the efficiency of the TRL-HPO process, it has yet to converge in a short time. The bottleneck in the training process originates from two main sources:

1. The training of individual models
2. Optimization rounds of the RL approach

The first source is a necessary evil because of the exploration requirements and the need to evaluate the actor's actions. Determining the training data size, the number of epochs, and learning rates are required to achieve close to optimal accuracy. During the exploration and action assessment phases, repetitive models can be generated. Storing these models can avoid retraining them, which

Challenge	Summary	Solutions
Computational Time	• Training of individual models • Optimization rounds of actor-critic approach	Storing the trained models in a hash table to avoid re-training
Exploration Strategy	DDPG's lack of direction	Application of RL methods with inherent randomness such as soft actor-critic approach
Reward Function	• Training many models to infer their contribution to classification results • Prioritization of accuracy	• Generating a block of similar layers • Inclusion of inference time, energy consumption, and number of operations per second
Reward Function	• Training many models to infer their contribution to classification results • Prioritization of accuracy	• Generating a block of similar layers • Inclusion of inference time, energy consumption, and number of operations per second
Generalizability, Scalability, and Transferability	• Scalability and Transferability concerns with larger datasets and complex architectures. • Generalizability to different IoT Environments	• Expansion of the input sequence and actor's block generation • Changing the mappings of the actor's output to adapt to different DNN models
Integration into Automated ML	N/A	Replacing the HPO process for IoT deployment frameworks

TABLE 3. Summary of challenges.

can overwhelm the available GPU or RAM resources, especially with a large ER buffer. An alternative method is storing already trained models in a database or a hash table, whereby the model is represented via its HP or its hash value. A salient issue relates to querying algorithms that should search for HP representations available in the database. The querying time will increase with the expansion of the database, requiring a more intricate search procedure. These methods should be investigated to budget the computational time towards more fruitful procedures that benefit RL-based solutions.

Exploration

RL implementations depend on the trial-and-error procedure in the exploration phase. However, a shortcoming of this phase, either through DDPG's random noise or ε-greedy algorithms [6], relates to its lack of direction. This means that the agent continues exploring unpromising areas of the search space with relatively poor performances. Therefore, it is imperative to include priors or rectify the exploration stage to reduce unnecessary exploration. Aspects of the BO process should be integrated into the RL-based strategies to streamline the exploration process. The randomness can be incorporated using a stochastic policy, which is part of the soft actor-critic policy gradient approach [15]. In the exploration stage, the stochastic policy outputs the normal distribution's mean and standard deviation. The standard deviation is progressively reduced to achieve a deterministic policy based on the reward function. This suggestion opens new frontiers toward examining methods that can associate the uncertainty of the stochastic policy with steering the RL's agent exploration.

Reward Function

The progressive reward function of TRL-HPO is foundational to gaining insights into the contribution of each layer to performance enhancement. However, this layer accumulation process requires training many models to infer their performance, restricting the model generation process to a few layers. These constraints can be sidestepped by generating similar layers at once, which is referred to as a block generation procedure. As such, computational time is gained at the expense of transparency and layer diversity. In resource-constrained environments, factors such as energy consumption, inference time, and number of computational operations per second can gain precedence over accuracy. These factors can be added to the reward function and assigned a weight based on application requirements.

Generalizability, Scalability, and Transferability

With each contribution toward the HPO procedure, questions of generalizability, scalability and transferability loom to undermine their usefulness; a concern that applies to RL-based solutions. In TRL-HPO, the actor and the critic generate and evaluate models using FCL, CNN, and MaxPool layers, necessary to build DNN models. When applied to complex datasets and architectures, skip and residual connections can be integrated into the TRL-HPO architecture. Towards that end, two main approaches can be followed:

1. Expand the input space to include more layers, which is reflected in the sequence length variable
2. Change the definition of the actor's output from the layer and its HPs to include the number of similar layers and their HPs.

Adopting any of these approaches should consider the trade-off between the availability of computational resources and the layer generation process's transparency. The same adaptation procedure can be followed when confronting a transferability concern.

The IoT environment is integrated into various fields, including smart cities, healthcare, buildings, and electric grids. This environment suffers from computational and communication resource scarcity, which can be detrimental to ML applications, such as anomaly detection, object detection, and forecasting. In connection to the IoT environment limitations, the TRL-HPO framework offers solutions manifested in three aspects:

1. Its generalizability to different types of DNN models, including CNN, Long short-term memory (LSTM) and FFNN by changing the mappings of the actor's output
2. Its low computational, storage, and processing footprint, compared to SOTA approaches

3. Its transparency that enables the reliance on part and not all of the CNN models and exchanging model parameters when the need arises.

Integration into AutoML

The HPO process is an important procedure in the AutoML pipeline and its enhancement is central to the wide-scale deployment and adoption of these pipelines. The TRL-HPO process is envisioned to replace the generic HPO processes that are part of the ubiquitous cloud computing ML deployment modules. These modules enhance the efficiency and accessibility of ML deployment, by providing user-friendly deployment strategies. TRL-HPO that outperformed SOTA approaches in its convergence time while maintaining good accuracy results benefits experts who have budgetary constraints and desire to prove the viability of their products/models for customers/investors alike.

Conclusion

The HPO process is a fundamental step in the ML pipeline that enhances model performance. However, its computational footprint is prohibitive for widespread adoption and current methods overlook the transparency in the layer generation process. These factors are imperative to realize IoT applications' function, AVs in particular, requiring model partitioning to fulfill the resource constraints of edge environments. To address these limitations, this article proposes TRL-HPO framework that combines transformers with an RL actor-critic approach. The attention mechanism, parallelization, and the progressive generation of layers are all novel properties of this framework within the transparency and time requirements. These advantageous factors were empirically established compared to SOTA approaches using the MNIST dataset. Moreover, a list of research questions scrutinizing RL-based solutions, including TRL-HPO, is presented with corresponding recommendations. Future work will target the presented questions to design a general-purpose HPO process.

Acknowledgement

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) and Compute/Calcul Canada.

References

[1] I. Shaer and A. Shami, "Corrfl: correlation-based neural network architecture for unavailability concerns in a heterogeneous iot environment," IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 1543–1557, 2023.

[2] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," The Journal of Machine Learning Research, vol. 20, no. 1, pp. 1997–2017, 2019.

[3] D. Baymurzina, E. Golikov, and M. Burtsev, "A Review of Neural Architecture Search," *Neurocomputing*, vol. 474, 2022, pp. 82–93.

[4] X. Xu *et al.*, "CNN Partitioning and Offloading for Vehicular Edge Networks in web3," *IEEE Commun. Mag.*, 2023.

[5] B. Zoph and Q. Le, "Neural Architecture Search with Reinforcement Learning," *Int'l. Conf. Learning Representations*, 2016.

[6] B. Baker *et al.*, "Designing Neural Network Architectures Using Reinforcement Learning," *Int'l. Conf. Learning Representations*, 2016.

[7] A. Vaswani *et al.*, "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," *Neurocomputing*, vol. 415, 2020, pp. 295–316.

[9] P. Neary, "Automatic Hyperparameter Tuning in Deep Convolutional Neural Networks Using Asynchronous Reinforcement Learning," *2018 IEEE Int'l. Conf. Cognitive Computing (ICCC)*, IEEE Computer Society, 2018, pp. 73–77.

[10] A. Iranfar, M. Zapater, and D. Atienza, "Multiagent Reinforcement Learning for Hyperparameter Optimization of Convolutional Neural Networks," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 4, 2021, pp. 1034–47.

[11] C.-H. Hsu *et al.*, "MONAS: Multiobjective Neural Architecture Search Using Reinforcement Learning," arXiv preprint arXiv:1806.10332, 2018.

[12] M. Tan *et al.*, "Mnasnet: Platform-Aware Neural Architecture Search for Mobile," *Proc. IEEE/CVF Conf. Computer vision and Pattern Recognition*, 2019, pp. 2820–28.

[13] T. P. Lillicrap *et al.*, "Continuous Control with Deep Reinforcement Learning," arXiv preprint arXiv:1509.02971, 2015.

[14] Y. LeCun, "The MNIST Database of Handwritten Digits," http://yann. lecun. com/exdb/mnist/, 1998.

[15] T. Haarnoja *et al.*, "Soft Actor-Critic Algorithms and Applications," arXiv preprint arXiv:1812.05905, 2018.

Biogrpahies

Ibrahim Shaer (ishaer@uwo.ca) received his B.S. degree in Computer Science from the American University of Beirut, Beirut, Lebanon in 2017 and his M.E.Sc. degree in Electrical and Computer Engineering from the University of Western Ontario, London, Canada in 2020. He is pursuing his Ph.D. in Electrical and Computer Engineering from the University of Western Ontario, London, Canada as part of the Optimized Computing and Communication Laboratory. His research interests include applications of Machine Learning in industrial buildings, such as the optimization of Heating, Ventilation, and Air Conditioning systems and anomaly detection and Machine Learning interpretation in high-dimensional spaces. He is an active IEEE volunteer and a member of the IEEE Computer Society.

Soodeh Nikan (snikan@uwo.ca) received a Ph.D. degree in Electrical and Computer Engineering from the University of Windsor in 2014. She is currently an Assistant Professor in software engineering with the Department of ECE, Western University, Canada. Her research is committed to artificial intelligence and machine learning, computer vision, data analytics, and signal processing. She has made significant contributions to optimized deep/machine learning-based technologies for highly demanding and safety-critical areas. She has an extensive academic and industry portfolio in AI and automotive research through her research on autonomous driving at Ford Motor Company and Western University. She holds a deep connection to the academic community, by her role as a Counselor for the IEEE London Ontario Section Branch and active participation in the Technical Program Committee of IEEE-sponsored events and reviewer for distinguished journals in electrical and computer engineering..

Abdallah Shami (abdallah.shami@uwo.ca) is a Professor in the Electrical and Computer Engineering Department, Western University, London, ON, Canada, where he is also the Director of the Optimized Computing and Communications Laboratory. Dr. Shami is a fellow of the Canadian Academy of Engineering (CAE). He has chaired key symposia for the IEEE GLOBECOM, IEEE International Conference on Communications, and IEEE International Conference on Computing, Networking and Communications. He was the elected Chair of the IEEE Communications Society Technical Committee on Communications Software and the IEEE London Ontario Section Chair. He is currently an Associate Editor of the *IEEE Transactions on Information Forensics and Security* and *IEEE Communications Surveys and Tutorials* journals.

IEEE
INTERNET OF THINGS
MAGAZINE

CALL FOR PAPERS

EMPOWERING PEOPLE WITH DISABILITIES USING IOT AND HUMAN-CENTERED TECHNOLOGIES

BACKGROUND

The World Health Organization estimates that 1.3 billion people, which represents about 16% of the global population, are presently living with some form of disability. Sadly, these individuals are also at higher risk of discrimination, poverty, social exclusion, violence, and abuse. Thankfully, global initiatives are rising to this challenge. The United Nations has set a target to foster social, economic, and political inclusion for all by 2030, regardless of factors such as disability. Similarly, the European Commission's Strategy for the Rights of Persons with Disabilities 2021-2030 seeks to achieve the full participation and non-discrimination of people with disabilities in society.

To face these challenges and ensure a welcoming and accessible society for everyone, investing in human-centered technologies is paramount. The Internet of Things (IoT) offers transformative potential for empowering individuals with disabilities. Technologies designed with the active input from IoT sensors and engagement of disabled communities can lead to the creation of Human-centered solutions that are inclusive and adaptive to various needs. For instance, IoT can enhance physical environments to be more accessible through smart home systems that allow for voice-activated or remotely controlled management of appliances, doors, and lighting, reducing physical barriers. Wearable IoT devices can also aid in real-time health monitoring and emergency response, ensuring that people with disabilities receive timely medical attention when necessary. Furthermore, IoT can facilitate better communication and interaction with technology via customized interfaces that adapt to the physical and sensory capabilities of users, such as gesture recognition systems for those unable to use traditional input devices.

Through a human-centered strategy and investments in IoT devices within everyday objects and environments, we can significantly enhance the independence and quality of life for people with disabilities, creating a more inclusive society. This approach unlocks innovative solutions for collaborative environments, empowering individuals with disabilities to promote their rights and opportunities. Ultimately, by embracing this transformative approach, we tap into the varied talents and contributions of every member, leading to increased stability, balance, and sustainability for all.

This Special Issue (SI) aims to bring together researchers, policymakers, and industry professionals to contribute their ideas and work on the topic of empowering people with disabilities through IoT and human-centered technologies. Subtopics of interest include, but are not limited to:

1. Sensor technology and computer processing algorithms, to improve safety, mobility, communication, and activities of daily living, etc.
2. Tracking technologies and wearable IoT devices, to enhance data collection processes and improve the quality of life for individuals with disabilities.
3. Accessible human-machine interfaces, featuring voice-controlled interfaces, alternative control devices, and adaptive user interfaces that learn and adapt to individual needs.
4. Studies on the social implications of implementing IoT and human-centered technologies for individuals with disabilities.
5. Real-world case studies and applications of human-centered communication technologies.
6. Cloud-based technologies for telehealth, remote care, collaborative learning, and working platforms.
7. Human-robot collaboration into healthcare, education, and workplaces for assistive and caregiving roles.
8. Use digital twins to simulate accessibility in the design of buildings, spaces, and products.

Researchers are invited to publish their experimental and theoretical results in detail. Submissions should align with the outlined topics and seek to spark meaningful discussions on the role of technology in fostering a more inclusive society.

SUBMISSION GUIDELINES

Manuscripts should conform to the *IEEE Internet of Things Magazine* standard format as indicated in the Information for Authors section of the Article Submission Guidelines.

All manuscripts to be considered for publication must be submitted by the deadline through the magazine's IEEE Author Portal site. Select the appropriate issue date and topic from the "Please Select an Article Type" drop-down menu.

IMPORTANT DATES

Submission Deadline
15 December 2024

Decision Notification
28 February 2025

Final Manuscript Due
7 March 2025

Publication Date
May 2025

GUEST EDITORS

Nagham Saeed (Lead Guest Editor)
School of Computing and Engineering, University of West London, UK
Nagham.Saeed@uwl.ac.uk

Carolina Lagartinho Oliveira
NOVA School of Science and Technology, NOVA University Lisbon, Portugal
ci.oliveira@campus.fct.unl.pt

Abdellah Chehri
Royal Military College of Canada, Canada
achehri@gmail.com

Gwanggil Jeon
Incheon National University, South Korea
gjeon@inu.ac.kr

IoT-Based Piano Playing Robot

Hsing-Hsin Huang and Yi-Bing Lin

Abstract

This article proposes PianoTalk, an IoT-based piano-playing robot that strives for excellence in construction, control methods, and appearance. The robot features two 19-finger piano-playing hands, each functioning as an independent IoT device controlled by an IoT development platform called IoTtalk. We propose six essential rules to enable a robot to simulate human piano playing. The objective extends beyond technical prowess to embody a thematic demonstration of musical capability. This innovation offers two key contributions. Firstly, it marks the first IoT-based piano-playing robot where each robot hand and sustain pedal functions as an autonomous IoT device. This enables cascading and synchronization of multiple robot hands for any off-the-shelf piano, overcoming the fixed hand limitation in previous solutions. For example, PianoTalk can assign either one or two robot hands to a 54-key piano and up to four robot hands for a 97-key piano duet. Secondly, PianoTalk leverages the advantages of IoT, enabling piano robots to perform orchestral pieces simultaneously in various locations. Connected to IoTtalk, it enables remote collaborations with other IoT-based musical instruments, such as the developed violin robot and smart gloves, for a comprehensive musical ensemble.

Introduction

The piano, originating in 17th-century Europe, has evolved into a versatile and expressive instrument often dubbed the "king of instruments." Its tonal range spans deep bass to bright treble, facilitating interpretation across various musical genres. From classical to popular music, the piano serves as both a solo and accompanying instrument, showcasing its adaptability.

The production of piano sound involves keys activating hammers that strike strings, resulting in a wide range of expressive capabilities. The first piano had only 54 keys. As piano music evolved, the keyboard range gradually expanded to meet the demands of composers, seeking broader expressive potential. By the 1890s, the modern piano had developed into an 88-key instrument. Today, some piano manufacturers customize pianos with 97 keys to meet specific needs, adding 9 keys in the lower register. These additional keys enhance resonance, producing a richer sound when playing other keys. Electronic keyboards or MIDI keyboards typically have 61 keys (with MIDI keyboards having even fewer).

Modern pianos, with 88 keys, cover seven octaves, making them the broadest among keyboard instruments. Electronic pianos use sensors and electronic components to replicate the traditional striking mechanism, enhancing playing effects.

Timbre control through dynamics, key touch, and pedals allows pianists to convey a range of emotions. Achieving excellence demands coordinated use of arms, wrists, and fingers, emphasizing the continuous dedication and practice required for high-level piano performance.

Based on an Internet of Things (IoT) platform called IoTtalk [1], this article introduces PianoTalk, an IoT-based piano-playing robot. Its objective is not only to showcase excellence in construction, control methods, and appearance but also to embody a clear thematic demonstration. PianoTalk features multiple 19-finger piano-playing robot hands, each functioning as an IoT device controlled by IoTtalk. Both the left and right hands aim to minimize lateral movement of the piano-playing mechanism and enhance the flexibility of the performance. We demonstrate that, overall, the piano-playing robot exhibits excellent performance. With well-controlled key positions, precise operation timing, and appropriate force, it is capable of producing beautiful music. In this article, we first propose six design rules for piano robots to evaluate existing approaches. Next, we delve into the PianoTalk architecture, outlining how the robot hands are implemented as IoT devices. Finally, we discuss the contributions of our solution and identify future research directions.

Rules for Piano Robot Design

Drawing from our decade-long experience in studying piano robots, we propose six essential rules to enable a robot to simulate human piano playing:

Rule 1 (Finger Technique): The mechanical robot must have control over the force exerted by its fingers to produce different volumes. It should accurately press the piano keys, with fast response times for pressing, releasing, and movement to avoid interference between fingers. Appropriate actuators like pneumatic cylinders are recommended for a quick response, and precise finger positioning control, such as using motors for linear actuators, is necessary for accurate key pressing.

Rule 2 (Hand Coordination): For a balanced performance, both hands of the piano-playing robot should collaborate, independently controlling the force of each hand. Typically, the right hand handles the main melody, while the left hand manages accompaniment. The robot should have at least two palms, each with multiple fingers, capable of independent and cooperative control.

Hsing-Hsin Huang is with Chung Yuan Christian University (CYCU), Taiwan;
Yi-Bing Lin (corresponding author) is with National Yang Ming Chiao Tung University (NYCU), Taiwan.

Digital Object Identifier: 10.1109/IOTM.001.2400017

Rule 3 (Musical Expression): Aligning with the rhythm of the musical score, the robot should convey emotions through controlling speed, intensity, and pedal usage during performance. It must be capable of rhythm control, regulating key force, and controlling the sustain pedal to enhance musical continuity.

Rule 4 (Playing Posture): Design the positions of both hands and fingers based on the piano keyboard arrangement, aiming for effective playing and aesthetic characteristics. The finger arrangement should conform to the keyboard design to effectively span, for example, 10 white keys.

Rule 5 (Music Interpretation): The one writing control commands must be able to understand musical notation and express musical intent. When the robot receives control commands based on sheet music, it should execute dynamic control according to the indications on the sheet music and accurately grasp the rhythm, fully expressing the beauty of the music.

Rule 6 (Remote Control): With the increasing maturity of IoT technology, the robot can utilize IoT techniques for control. This includes the possibility of remotely controlling the operations of the orchestra.

The force applied to press the piano keys is typically designed to be light, making it easy for humans to press the keys to their maximum depth (e.g., H = 12 mm; see (1) in Fig. 1a). When played by a robot, sound is triggered only if two conditions are met: the piano key must reach a specific effective position (e.g., 7mm), and the force of the play must exceed a preset minimum value (e.g., 0.2gf). The speed of play often influences the motion of the piano keys and can be used to determine the intensity of the play.

The human palm, with its multiple degrees of freedom, forms a three-dimensional structure resembling a cup when flexed during play. Each finger can rotate, enabling another finger to reach the key's position in advance, eliminating delays. Fig. 1b illustrates rotation and position changes of the thumb and middle finger, accompanied by wrist movement.

Clearly, it is difficult for a robot to simulate such complicated human posture. When using a robot to play the piano, we take advantage of fast one-dimensional movement characteristics of the robot. After the control system receives the musical instructions, it determines the target position for each note and calculates the delay t_p ((2) in Fig. 1a) to press the keys based on the rhythm. To address delay t_L for hand movement and gaps before and after key presses ((3) in Fig. 1a), the piano's sustain mechanism (sustain or legato pedal) is commonly utilized to eliminate the phenomenon of musical notes pausing.

Related Work

The study in [2] introduced ROBOPIANIST, a benchmarking suite for high-dimensional control, focused on testing precision, coordination, and planning for an animated piano robot. It offers a versatile benchmarking environment for quantitative assessment and research avenues in multi-task learning, zero-shot generalization, and multimodal learning. We notice that [2] and specific existing approaches [3–5], when finely tuned, can attain remarkable performance levels in certain aspects. However, these studies focus solely on animated simulation without actual implementation of physical robotic hands. They overlook the real mechanical motion characteristics and limitations. Especially during high-speed piano playing, although the simulated mechanisms quickly move to key positions, the actual robot mechanisms struggle to avoid acceleration, deceleration, and collision issues. Therefore, these studies do not adhere to the finger technique principles (**Rule 1**) applicable to piano playing.

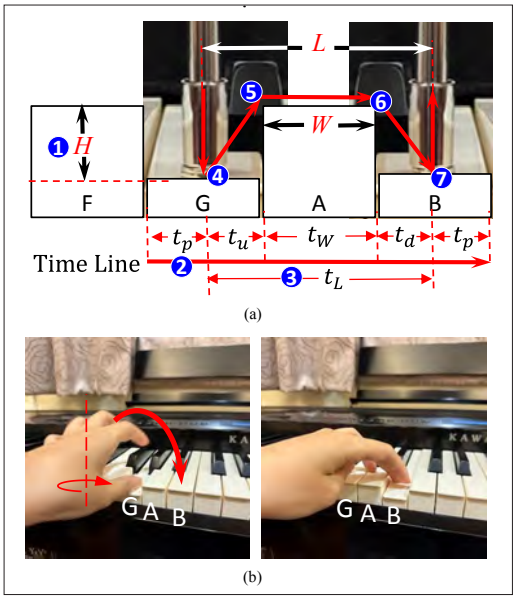


FIGURE 1. Piano playing posture: a) the timing diagram for finger movement; b) human palm posture.

Deep learning, renowned for robust computational tools, finds diverse applications in data and signal processing. Its increasing focus on music signal processing, explored in recent works like Moysis et al.'s review [6], highlights music information retrieval and music generation. The review identifies emerging directions for future research in this dynamic intersection of deep learning and music processing. Although the review does not delve into the technology of piano robots, the application of artificial intelligence to generate music in various styles not only represents a potential future direction for piano robot development but also aligns with the principles of musical interpretation (**Rule 5**).

Some innovative approaches [7, 8] tackle limitations found in flexible smart gloves, often associated with processing complexity and stability issues during mass production. An integrated full-print production flexible smart glove utilizes a flexible printed circuit board with topological carbon-silver strain sensors and employs deep learning for motion intention prediction. This allows the prediction of finger motion intentions and proactive response, reducing communication latency and enhancing remote interaction. These techniques can be applied in PianoTalk to improve the implementation of **Rule 6**.

The study in [9] developed the Supernumerary Robotic 3rd Thumb with two degrees-of-freedom, controlled by the user's body, augmenting humans with an extra thumb. Pianists learned to play with 11 fingers within an hour. Evaluation

FIGURE 2. The PianoTalk hardware architecture.

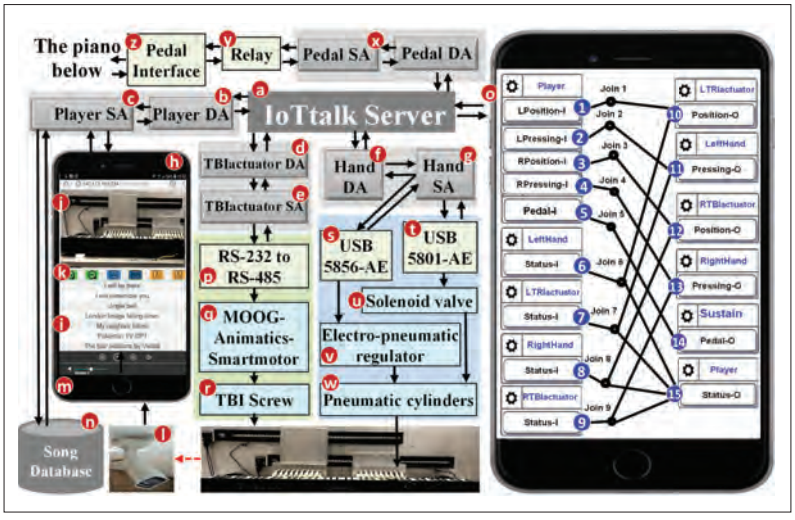


FIGURE 3. The PianoTalk software architecture.

of naive and experienced players showed augmented performance, explained by the human augmentation motor coordination assessment, demonstrating how supernumerary robotics enhance human skills based on individual motor coordination. This study references additional fingers to assist and enhance the user's piano-playing abilities. However, in the design of piano robots, achieving superior performance can be easily accomplished by simply adding more fingers to a single palm as we do in PianoTalk.

We were pioneers in the development of professional piano robots [10]. In 2010, we introduced the design of an electronic piano-playing robot equipped with five-fingered hands, each possessing two degrees of freedom. The robot utilizes stepping motors, pneumatic cylinders, and a specially-designed controller to play electronic piano based on user-defined sheet music. Linear motors assist in reaching keys, and a Digital I/O PCI card sends commands to the controller. Building upon the foundation of the six essential rules for a piano robot, we have developed PianoTalk to significantly advance our previous work, integrating a stable finger robot design and incorporating IoT extension.

In [11], the authors explored enhancing robotic manipulators' adaptability and dexterity using a multi-material 3D-printed anthropomorphic skeleton hand. The 'Conditional Model' introduced anisotropic stiffness in soft-rigid hybrid systems, enabling diverse dynamic interactions. This approach allowed a single skeleton hand to play varied piano music, demonstrating superior adaptability over rigid end effectors. Meanwhile, [12] delved into the shadow dexterous hand, a tendon-driven robotic hand, in the context of piano playing. The study analyzed hand postures, force application for volume control, and timing challenges. A software pipeline incorporating inverse kinematics, trajectory scaling, and Mixture Density Networks compared robotic and human hand performance in key-pressing tasks. Despite advancements, both studies face a challenge: the transient response speed of flexible mechanisms is known to be slower than rigid mechanisms. The inherent deformability of flexible mechanisms requires a longer time to stabilize after external forces, impacting finger dexterity principles (**Rule 1**).

None of the studies mentioned above adhere to **Rule 6** and are unable to harness the advantages of IoT for the simultaneous performance of robot musical instruments in different locations.

PianoTalk Architecture

In accordance with the 88-key piano keyboard, Fig. 2 illustrates the PianoTalk hardware architecture, showcasing dimensions of 1500 cm in length, 650 cm in width, and 1602 cm in height. In this example, an electronic keyboard (Fig. 2 (1)) is played by two 19-finger robot hands (Fig. 2 (2) and (3)). The robot hands are arranged in parallel, imitating human performance, with the right hand playing the main melody and the left hand handling the accompaniment.

The movements of the robot hands are controlled by two precision ball screws, referred to as TBI Single Axis Actuators (abbreviated as TBI actuators; see Fig. 2 (4) and (5)) [13]. The robot fingers' key hits are triggered by direct-ported solenoid valves for remote control base (SRVB; see Fig. 2 (6) and (7)) and the CVTR series electro-pneumatic regulators (Fig. 2 (8) and (9)) [14]. The NC series FRL (Filter, Regulator, and Lubricator) Combinations (Fig. 2 (10)) are used to filter compressed gas, regulate pressure, and provide lubrication between pneumatic components. A touch panel (Fig. 2 (11)) is used to control and show the status of PianoTalk.

In PianoTalk, the TBI actuators and the robot hands are implemented as IoT devices, and their software modules are managed by an IoT platform called IoTtalk [1]. Within this platform, each IoT device is equipped with two software modules. The Sensor and Actuator Application (SA) is responsible for the intelligence of the IoT device, while the Device Application (DA) is responsible for communication with other IoT devices through the IoTtalk server (Fig. 3a). PianoTalk incorporates three types of IoT devices, and their software modules are depicted in the gray areas (Fig. 3b–g and 3x).

The Player software (Fig. 3b and 3c) implements a control board that enables users to instruct PianoTalk to play a song through a web-

based browser, e.g., on a smartphone (Fig. 3h) or the control panel (Fig. 2 (11)). Users can select a song to play (Fig. 3i) and remotely observe piano performance through the view window (Fig. 3j), with the ability to zoom and rotate the camera (Fig. 3k). The camera (Fig. 3l) is equipped with a microphone, allowing users to control the volume of the piano sound (Fig. 3m). Our IoT-based design satisfies **Rule 6**.

When the user selects a song, the Player device retrieves the "playing format" of the song from the database (Fig. 3n). The playing format translates sheet music into instructions, specifying the positions of TBI actuators, the strengths of finger-pressing, and the control of the sustain pedal. The details of the playing format will be elaborated later. These IoT devices communicate through "device features" (DFs) defined in the DA. An IoT device uses input DFs (IDFs) to send messages to the IoTtalk server. The IoTtalk server then forwards the messages to the output DFs (ODFs) of the target IoT device.

PianoTalk offers a web-based graphical user interface (GUI; Fig. 3o) to configure the DF communications. In this GUI, every IoT device is represented by two "device" icons—one for input placed on the left side of the window and one for output on the right side. Within a device icon, there are several small icons representing DFs. The IDF name is appended with "–I" and the ODF name is appended with "–O." By simply dragging a Join line from an IDF icon to an ODF icon, the IDF can send messages to the ODF.

In the input device icon "Player," there are four IDFs: "LPosition-I" (the position for the left hand; see Fig. 3 (1)), "LPressing-I" (the pressing strengths for the 19 fingers in the left hand; see Fig. 3 (2)), "RPosition-I" and "RPressing-I" for the right hand (Fig. 3 (3) and (4)), and Pedal-I" (on/off for the sustain pedal; see Fig. 3 (5)). In the output device icon "Player," there is one ODF: "Status-O" (Fig. 3 (15)), which receives the playing result of the TBI actuators and robot hands. By dragging a Join line from an IDF icon to an ODF icon, the IDF can send messages to the ODF. For instance, using "Join 1" ((1) → (10) in Fig. 3), the Player sends the position coordinate to the left TBI actuator.

When a song is selected for playback (Fig. 3i), the Player IoT device (Fig. 3c) retrieves the song's bit streams from the Song Database (Fig. 3n), as further detailed in Fig. 5. The Player device then partitions each bit stream into segments for the playing IoT devices, including the TBI actuators, robot hands, and pedals, facilitated by Joins 1-5. Subsequently, the playing devices synchronize with each other for each bit stream and report their statuses to the Player via Joins 6-9. This process repeats until all bit streams are completely executed.

TBI Actuator IoT Device

The software modules of a TBI actuator are illustrated in Fig. 3d and 3e. The TBI actuator SA controls the hardware through the Advantech DAQNavi LabVIEW. Specifically, the RS232 format is translated to RS486 format (Fig. 3p) to invoke the motor (Fig. 3q) to drive the ball screw (Fig. 3r). We use precision-grade, high-load single-axis actuators [13] with a core composed of a precisely ground ball screw (Fig. 4 (1)), achieving a positioning accuracy of 0.04mm, making it ideal for driving the piano

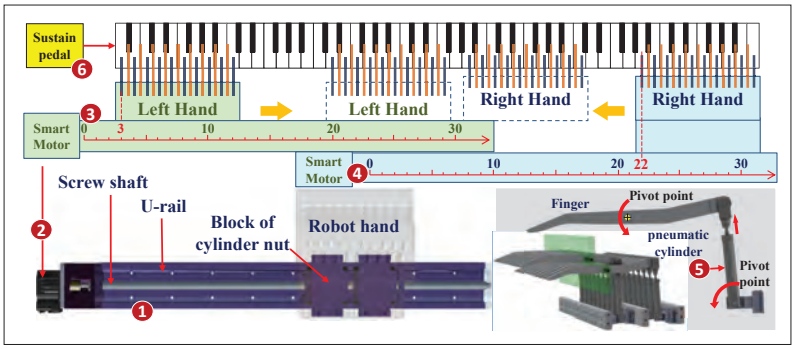


FIGURE 4. The TBI actuator and the hand robot.

robot hands. The TBI single-axis actuator is driven by a MOOG Animatics Smartmotor (Fig. 4 (2)), capable of linear motion only along a single axis. Two Smartmotors (Fig. 4 (3) and (4)) are used to move the left and the right hands.

The TBI actuators are fixed in the platform, and each of them has its own coordinate system (Fig. 4 (3) for the left hand and (4) for the right hand). The coordinates are spaced at the width of the white keys, and the origin of the coordinate coincides with the origin of the motor. For ease of music notation, we set the starting point of the left hand coordinate to 3 and the starting point of the right hand coordinate to 22 for the 88-key piano. Consider the TBI actuator for the left hand. Its SA operates with the following steps:

Step T1. Prior to commencing the performance, the SA positions the left hand to its starting location, i.e., position 3 (Fig. 4 (3)).

Step T2. Upon receiving the song's playing format from the Player through Join 1 (Fig. 3 (1) → (10)), the SA instructs the Smartmotor to move the left hand to the initial position of the musical note, commencing the performance thereafter. Specifically, the SA interprets the left hand's positioning coordinate into Smartmotor's rotation angle, transmits the rotation angle to the Smartmotor through RS232/RS485, guiding the ball screw to the designated keyboard position [13].

Step T3. After the Smartmotor stops rotating, the hand has moved to position (4) in Fig. 1a. Subsequently, the SA instructs the left hand to activate the fingers through Join 7 ((7) → (11) in Fig. 3) and reports the result to the Player ((7) → (15) in Fig. 3).

Step T4. If the last notes of the song have been played, go to Step T1. Otherwise, go to Step T2 for playing the next notes.

The TBI actuator's SA is tasked with calculating the speed of the smart motor, ensuring that the robot fingers can press the piano keys at the right time. Referring to Fig. 1a, when a single finger changes its position on the piano, moving from the "G" key to the "B" key — excluding the time for pressing the "G" key and releasing the "B" key — the interval delay time is determined by $t_L = L/V_x$ ((3) in Fig. 1a), where V_x is the average horizontal movement speed of the finger. To avoid touching the "A" key, the finger also needs vertical movement. This delay can be divided into three segments, where t_u is the delay for vertical up movement the finger ((4) → (5) in Fig. 1a), t_W is the delay for horizontal movement ((5) → (6) in Fig. 1a), and t_d is the delay for vertical down

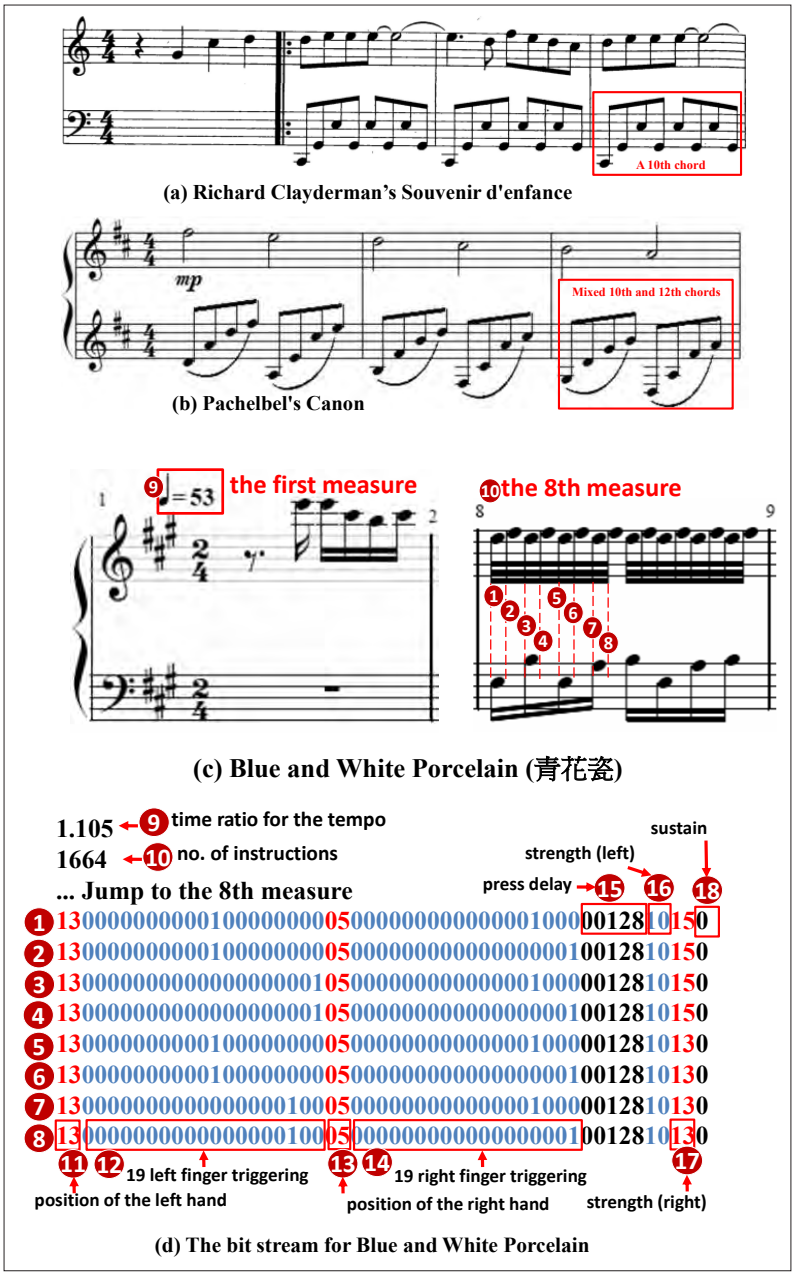


FIGURE 5. The PianoTalk playing format.

movement of the finger ((6) → (7) in Fig. 1a). Then $t_L = t_d + t_W + t_u$. Our measurements show that $t_d \cong 10$ms, $t_u \cong 15$ms, $t_W \cong 100k$ ms for k keys. The PianoTalk successfully coordinates the hand movements through Joins 1–2 and Joins 3–4 of the IoTtalk configuration in Fig. 3. Therefore, **Rule 2** is satisfied.

Robot Hand and Sustain Pedal IoT Devices

The software modules of a robot hand are illustrated in Fig. 3f and 3g. When designing the robot hand, the positions of the palm and fingers can be determined relative to the music's difficulty. Simultaneously, by adhering to common accompaniment chord fingerings, like major and minor triads, 10th and 12th chords (Fig. 5a and 5b), the fingers in a palm should be configured to span 10 or 12 white keys. This design minimizes the frequency of hand movement by increasing the number of fingers, thereby reducing control challenges. This compromise aligns with both **Rule 1** and **Rule 4**.

In PianoTalk, a palm has 19 fingers, with the first finger positioned as a reference point; odd-numbered fingers handle white keys, and even-numbered fingers handle black keys.

In the design of mechanical fingers and palm, previous studies [10, 12] employed flexible finger-joint mechanisms, which are more susceptible to deformation or vibration when exposed to external forces, necessitating a longer time to achieve a stable state. In contrast, PianoTalk enhances the rigidity design of the palm and fingers (Fig. 4 (5)), aligning with the playing principles of a piano-playing robot. Our rigid mechanism features a more robust structure, undergoing smaller vibration amplitudes when subjected to external forces, enabling them to rapidly attain a stable state. By applying the lever principle, a single pneumatic cylinder propels one finger with two pivot points, culminating in the creation of highly rigid fingers. Our design adheres to the principles of finger dexterity (**Rule 1**).

When using a robot to play the piano, it is essential to consider the finger's movement characteristics. As depicted in Fig. 1a, each musical note involves four stages for the robot to press the piano key: starting pressing the key (with the delay t_h), pressing the key (with the delay t_p), releasing the finger (with the delay t_h), moving the palm (with the delay t_W), and starting pressing the finger again. The timing for these four stages must align with the rhythm requirements.

After receiving musical information from the database (Fig. 3n), the SA of the Player determines each target position based on the notes and calculates the time interval t_p to press the piano key according to the rhythm. Subsequently, through the Pressing-O ODF (Fig. 3 (11)), the DA of the left hand robot receives the t_p delay and the strength values of pressing for each finger. In the current implementation, this data is pre-calculated and stored in the database in advance, eliminating the need for the Player's SA to perform real-time calculations.

To satisfy **Rule 3**, we need to eliminate delay t_L in hand movement. We design the TBI actuator to move quickly, with t_L being 100 k + 25ms for k keys, and the piano's sustain device is typically utilized to eliminate note pauses caused by t_L, which is controlled by a relay (Fig. 3y) and a pedal interface (Fig. 3z) in PianoTalk. The sustain mechanism is controlled by the Pedal SA (Fig. 3x), which receives the on/off instruction from the Player through Join 5 ((5) → (14) in Fig. 3).

As we mentioned before, to simplify control, we adopt a design with 19 fingers capable of spanning a 10-octave range. Each finger is driven by a pneumatic cylinder (Fig. 4 (5)), not only facilitating control but also accelerating the overall response speed [14]. To prevent finger collisions with the keys and minimize pneumatic cylinder noise, acrylic covers are placed around the palm. These covers not only eliminate noise but also enhance the aesthetic appearance. PianoTalk meets **Rule 1** through the above design. Consider the left robot hand as an example. The Hand SA works with the following steps.

Step R1. Prior to commencing the performance, the SA turns the pneumatic cylinders of all fingers "off."

Step R2. When the SA receives the song's playing format from the Player through Join 2 ((2) → (11) in Fig. 3) and the alert from the left TBI actuator through Join 7 ((7) → (11) in Fig. 3), it sends an analog signal to the electro-pneumatic regulator [14] via USB5856 ((s) → (v) in Fig. 3), with a variable range set from 0 to 10 volts. The regulator then regulates the air intake to the solenoid valves, determining the force applied to the keys.

Step R3. Simultaneously, the SA transmits on/off signals to the solenoid valves via USB5801 (Fig. 3t → u). Subsequently, the solenoid valves control the pneumatic cylinders (Fig. 3w) of the 19 fingers, enabling them to strike the desired keys.

Step R4. After all target fingers have pressed for the delay t_p, the SA releases the fingers. If the last notes of the song have been played, proceed to **Step R1**. Otherwise, the SA instructs the left TBI actuator to move to the next position via Join 6 ((6) → (10) in Fig. 3) and reports the result to the Player ((6) → (15) in Fig. 3). Then, proceed to **Step R2** to playg the next notes.

Playing Format

The Player SA (Fig. 3c) retrieves the playing format translated from the sheet music in the database (Fig. 3n). Figure 5c illustrates a portion of sheet music of a song "Blue and White Porcelain." According to the indication in the top left corner of the first bar of the score ((9) in Fig. 5c), the overall tempo of the music is 53 beats per minute (BPM), representing an actual time of 1132ms per beat.

The composition of the song spans 104 measures, featuring the briefest note introduced in the 8th measure — a series of 16 continuous 1/8 beat ornamentations ((10) in Fig. 5c). Consequently, in the composition instructions for PianoTalk, each measure is divided into 16 notes, totaling 1664 instructions to be issued from the Player's SA.

In PianoTalk, the default delay for each beat is 1024 ms. After dividing into 1/8 beats, the default time interval between two instructions is 128 ms. Given that the actual duration of one beat in "Blue and White Porcelain" is 1132 ms, the tempo ratio of 1132/1024 = 1.105 represents the proportion (excluding sound pauses) used to adjust the musical instructions to meet real-time requirements. Therefore, the actual instruction time interval is 128 × 1.105 = 141.4 ms. Modifying the ratio value for the tempo allows us to freely accelerate or decelerate the tempo of the performance, thus altering the rhythm.

The PianoTalk playing format of "Blue and White Porcelain" is the bit streams illustrated in Fig. 5d. The first line indicates the tempo ratio ((9) in Fig. 5d), while the second line indicates the total number of instructions ((10) in Fig. 5d). Each line from the third to the 1666th contains a bit stream of instructions that the Player SA will send to the TBI actuators, the robot hands, and the sustain pedal.

The playing bit-stream formats (1)–(8) in Fig. 5d correspond to the first part of the 8th measure ((1)-(8) in Fig. 5d). Every line has 8 fields.

Field 1 represents the position of the left hand ((11) in Fig. 5d).

Field 2 represents 19 left finger triggering ((12) in Fig. 5d) where 0 represents "off" and 1 represents "on."

Field 3 represents the position of the right hand ((13) in Fig. 5d).

Field 4 represents 19 right finger triggering ((14) in Fig. 5d).

Field 5 represents the press delay t_p ((15) in Fig. 5d) where 128 represents 141.4ms.

Fields 6 and 7 represent the finger strengths ((16) and (17) in Fig. 5d). The strength ranges from 0 to 20, which is mapped to 0 to 10 volts for the electro-pneumatic regulators) [14]. Note that in this case, the strength of the right hand gradually weakens from Lines (1)-(4) to Lines (5)-(8), decreasing from 15 to 13.

Field 8 represents the on/off of the sustain pedal ((18) in Fig. 5d).

Discussions and Conclusions

This article presents six essential rules for piano robot design. Drawing upon these rules, we have developed PianoTalk, utilizing a robust industrial IoT framework known as IoTtalk. PianoTalk serves as part of our efforts for IoTtalk to contribute to the advancement of the IoT industry.

PianoTalk makes two major contributions. First, PianoTalk is the first IoT-based piano-playing robot, where every robot hand, TBI actuator and sustain pedal is an independent IoT device. Therefore, for any off-the-shelf piano, multiple robot hands can be cascaded and synchronized by IoTtalk. In previous piano robot solutions, the number of robot hands is fixed (either 1 or 2). In PianoTalk, for example, we can allocate one or two robot hands to a 54-key piano. For a 97-key piano, up to four robot hands can be allocated for a piano duet performance.

Second, PianoTalk can leverage the advantages of IoT for playing robot musical instruments in different locations. In [15], we have developed the IoT-based violin robot, and the violin robot along with the smart gloves [8] are already connected to IoTtalk to perform remote performances with PianoTalk, satisfying **Rule 6**.

We have designed the playing format (Fig. 5d) capable of generating instructions for PianoTalk performances, meeting the requirements of Rules 3 and 5. In the current implementation, we still require experienced piano experts to generate these instructions. In the future, some directions for PianoTalk improvements include:

1. Volume Control: Enhance the robot's performance by controlling variations in volume, making the playing more human-like.
2. Ensemble Connection: Integrate the piano-playing robot with existing instrument-playing robots to create a robotic orchestra.
3. Automatic Transcription Program: Currently, manual transcription is time-consuming and error-prone. Establish an automatic transcription program for converting sheet music into control data, facilitating future development.
4. Key Position Calibration: Although the movements of MOOG-Animatics Smartmotors (Fig. 3 (2)) are accurate, and we seldom need to fine-tune for exact key positions, in the current PianoTalk implementation, fine-tuning is conducted manually. In the future, we will utilize images captured by the camera (Fig. 3j) to periodically and automatically adjust the key positions using an AI model, such as YOLO.

Drawing upon these rules, we have developed PianoTalk, utilizing a robust industrial IoT framework known as IoTtalk. PianoTalk serves as part of our efforts for IoTtalk to contribute to the advancement of the IoT industry.

Two of the best non-robot hand systems for automating piano playing are PianoDisc [16] and Spirio [17]. In the future, PianoTalk aims to achieve the performance level of PianoDisc and Spirio using robotic hands.

Demo Videos

Three demo songs: https://www.youtube.com/watch?v=dr4F6V3K3BM (Blue and White Porcelain), https://youtu.be/uLTq6ScpONY (Love Confession; including rhythm accompaniment), and https://youtu.be/bHVzBdrUtPE (Cannon).

Acknowledgment

This work was supported in part by National Science and Technology Council (NSTC) 112-2321-B-A49-018, NSTC 112-2221-E-A49-049, NSTC 112-2420-H-A49-001, NSTC 112-2634-F-A49-004, NSTC 112-2221-E-033-023, NCKU Miin Wu School of Computing, Quanta and NYCU/NCKU AI Joint Research Center, China Medical University Hospital, Research Center for Information Technology Innovation, Academia Sinica.

References

[1] Y. -B. Lin *et al.*, "IoTtalk: A Management Platform for Reconfigurable Sensor Devices," *IEEE Internet of Things J.*, vol. 4, no. 5, Oct. 2017., pp. 1552–62

[2] K. Zakka *et al.*, "RoboPianist: Dexterous Piano Playing with Deep Reinforcement Learning," *Conf. Robot Learning (CORL) 2023*, arXiv:2304.04150v1 [cs.RO] 9 Apr 2023.

[3] H. Wang *et al.*, "Data-Driven Simulation Framework for Expressive Piano Playing by Anthropomorphic Hand with Variable Passive Properties," *2022 IEEE 5th Int'l. Conf. Soft Robotics (RoboSoft)*, Edinburgh, United Kingdom, pp. 300-305, 2022

[4] H. Xu *et al.*, "Towards Learning to Play Piano with Dexterous Hands and Touch," *2022 IEEE/RSJ Int'l. Conf. Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 10,410–16.

[5] H. Wang *et al.*, "Reduced-Order Modeling of a Soft Anthropomorphic Finger for Piano Keystrokes," 2023, F. Lida *et al.* (Eds.), *Towards Autonomous Robotic Systems*, Lecture Notes in Computer Science, vol. 14136, Springer, Cham, 2023.

[6] L. Moysis, "Music Deep Learning: Deep Learning Methods for Music Signal Processing — A Review of the State-of-the-Art," *IEEE Access*, Feb. 2023.

[7] Y. Li *et al.*, "Fully Flexible Smart Gloves and Deep Learning Motion Intention Prediction for Ultralow Latency VR Interactions," *IEEE Sensors Letters*, vol. 7, no. 9, Sept. 2023.

[8] Y. -B. Lin, H. Luo and C. -C. Liao, "CATtalk: An IoT-Based Interactive Art Development Platform," *IEEE Access*, vol. 10, 2022, pp. 127,754–69.

[9] A. Shafti *et al.*, "Playing the Piano with A Robotic Third Thumb: Assessing Constraints of Human Augmentation, 2021, *Sci Rep 11*, 21375.

[10] J.-C. Lin *et al.*, "Electronic Piano Playing Robot," *Int'l. Symp. Computer, Communication, Control and Automation*, 2010.

[11] J. Hughes *et al.*, "An Anthropomorphic Soft Skeleton Hand Exploiting Conditional Models for Piano Playing," *Science Robotics*, vol. 3, p. 25.

[12] B. Scholz, "Playing Piano with a Shadow Dexterous Hand," MS thesis, Universität Hamburg, 2019.

[13] TBI Motion, Single Axis Actuator; accessed 2024: https://www.tbimotion.com.tw/en/category/single-axis-actuator.

[14] CHENLC, CVTR Series Electro-Pneumatic Regulator; Accessed 2024: https://www.chelic.com/document/TW/technical/QRuse/en/CVTR_all-E.pdf.

[15] H. -H. Huang and Y. -B. Lin, "ViolinTalk: Violin Robots as Internet of Things Devices," *IEEE Access*, vol. 11, 2023, pp. 23,846–61.

[16] Pianodisc; accessed 2024: https://pianodisc.com/what-is-pianodisc/.

[17] Spirio; accessed 2024: https://www.steinway.com/spirio.

Biographies

Hsing-Hsin Huang (hhh@cycu.edu.tw) received his Ph.D. degeee in mechanical engineering from Florida Institude of Technology, USA, in 1992. During 1992 - 2018, he joined Minghsin University of Science and Technology (MUST), and became Vice President of MUST in 2016. Huang joined Chung Yuan Christian University (CYCU) in 2018 as a professor in the Department of Mechanical Engineering. Huang's major is automation control. He has presided over a number of industry-university cooperation projects and holds dozens of patents. He is also the host of the Intelligent Automation Industry-Academic Technology Alliance of the National Science Council, committed to promoting the development of automation technology in the industry.

Yi-Bing Lin [M'96, SM'96, F'03] (yblin@ieee.org) is Winbond Chair Professor of National Yang Ming Chiao Tung University (NYCU), Chair Professor of College of Humanities and Sciences, China Medical University, Miin Wu School of Computing, National Cheng Kung University, Department of Computer Science and Information Engineering, Asia University, Research Center for Information Technology Innovation, Academia Sinica, and College of Artificial Intelligence, NYCU. He received his Bachelor's degree from National Cheng Kung University, Taiwan, in 1983, and his Ph.D. from the University of Washington, USA, in 1990. From 1990 to 1995 he was a Research Scientist with Bellcore (Telcordia). He then joined NYCU, where he remains. In 2010, Lin became a lifetime Chair Professor of NYCU, and in 2011, the Vice President of NYCU. During 2014 - 2016, Lin was Deputy Minister, Ministry of Science and Technology, Taiwan. Lin is an AAAS Fellow, ACM Fellow, and IET Fellow.

IEEE COMSOC TRAINING

ADVANCED TOPICS IN WIRELESS COURSE SERIES

Register for all three courses in one transaction and use promo code ATW2024F to save 15% on the registration fee.

5G RAN and Core Network: Architecture, Technology Enablers and Implementation Aspects
16–17 October // 10:30 am–3:30 pm EDT

O-RAN: Disrupting the Radio Access Network through Openness and Innovation
30–31 October 2024 // 10:30 am–3:30 pm EDT

Machine Type Communications in 5G and Beyond
13–14 November 2024 // 10:30 am–3:30 pm EST

Successfully complete all courses in the **Advanced Topics in Wireless** course series and earn a digital badge.

Interested in purchasing this training for your employees? Contact iwc-mktg@ieee.org for more details.

Online Courses ▪ www.comsoc.org/training

Deep Cooperation in ISAC System: Resource, Node and Infrastructure Perspectives

Zhiqing Wei, Haotian Liu, Zhiyong Feng, Huici Wu, Fan Liu, Qixun Zhang, and Yucong Du

Abstract

With the emerging Integrated Sensing and Communication (ISAC) technique, exploiting the mobile communication system with multi-domain resources, multiple network elements, and large-scale infrastructures to realize cooperative sensing is a crucial approach satisfying the requirements of high-accuracy and large-coverage sensing in Internet of Everything (IoE). In this article, the deep cooperation in ISAC system including three perspectives is investigated. In the microscopic perspective, namely, within a single node, the sensing information carried by time-frequency-space-code multi-domain resources is processed, such as phase compensation, coherent accumulation and other operations, thereby improving the sensing accuracy. In the mesoscopic perspective, the sensing accuracy could be improved through the cooperation of multiple nodes. We explore various multi-node cooperative sensing scenarios and present the corresponding challenges and future research trends. In the macroscopic perspective, the massive number of infrastructures from the same operator or different operators could perform cooperative sensing to extend the sensing coverage and improve the sensing continuity. We investigate network architecture, target tracking methods, and the digital twin assisted large-coverage cooperative sensing. Simulation results demonstrate the superiority of multi-node and multi-resource cooperative sensing over non-cooperative sensing. This article may provide a deep and comprehensive view on the cooperative sensing in ISAC system to enhance the performance of sensing, supporting the applications of IoE.

Introduction

Internet of Things (IoT), Artificial Intelligence (AI), and automation technologies are reconfiguring traditional industries, opening the era of Internet of Everything (IoE). The scenarios of IoE are transferring from pure human to the symbiosis of human, intelligent machines, and massive number of sensors. These applications urgently need to be supported by new information infrastructures with the integration of sensing and communication. With the development of Integrated Sensing and Communication (ISAC) technique [1], the mobile communication system, as the crucial infrastructure to support the emerging IoE, is constantly breaking through the pure communication function and integrating radar sensing function. Notably, the International Telecommunication Union (ITU) has identified ISAC as one of the scenarios of the Sixth-Generation (6G) mobile communication system [2].

The mobile communication system realizes radar sensing with ISAC technique by analyzing the echo of ISAC signal [3], thereby realizing target localization, environment reconstruction, etc. ISAC not only enhances the utilization of spectrum and hardware resources, but also realizes the coupling of digital and physical spaces and the mutual benefit of communication and sensing functions [1]. The Millimeter-wave (mmWave), Terahertz (THz), and massive Multiple Input Multiple Output (MIMO) in mobile communication system are developing rapidly, which guarantee the feasibility of ISAC technique [4]. However, in the applications of IoE, such as the sensing of vehicles and Unmanned Aerial Vehicles (UAVs), the fading and path occlusion of ISAC signals degrade the signal quality drastically, posing challenges to achieve high-accuracy, large-coverage, and continuous sensing. To this end, it is urgent to explore cooperative sensing approaches within the ISAC system to enhance sensing performance.

Cooperative sensing in ISAC system could be realized from the following three comprehensive perspectives according to the scope of cooperation.

- In the microscopic perspective, i.e., within a single node, there exists resource-level cooperation to improve sensing accuracy by fusing the sensing information carried in the time-frequency-space-code multi-domain resources.
- In the mesoscopic perspective, sensing accuracy could be improved through the cooperation of multiple nodes, including Base Station (BS), User Equipment (UE), UAV, and so forth.
- In the macroscopic perspective, the large number of infrastructures from one operator or even multiple operators could be applied to extend the sensing area and improve the continuity of sensing.

However, realizing the deep cooperation of ISAC system in the perspectives of resource, node and infrastructure, continues to face the following challenges.

- **Resource-level Cooperation:** The inconsistency of physical-layer parameters on fragmented multi-domain resources brings challenges

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62271081, in part by the National Key Research and Development Program of China under Grant 2020YFA0711302, and in part by the Fundamental Research Funds for the Central Universities under Grant 2024ZCJH01.

Zhiqing Wei (corresponding author), Haotian Liu (corresponding author), Zhiyong Feng, Huici Wu, Qixun Zhang, and Yucong Du are with Beijing University of Posts and Telecommunications, China; Fan Liu is with Southern University of Science and Technology, China.

Digital Object Identifier: 10.1109/IOTM.001.2400042

to sensing information fusion. For instance, differences in subcarrier spacing, carrier frequency and length of Orthogonal Frequency Division Multiplexing (OFDM) symbol between high and low frequency bands hinder the seamless fusion of sensing information collected from different frequency bands.

- **Node-level Cooperation:** The fusion of the sensing information from multiple nodes with non-synchronization or low-accuracy synchronization level in space, frequency and time domains is challenging.
- **Infrastructure-level Cooperation:** The rapid and seamless handover among multiple BSs from one operator or different operators is challenging.

Facing the above challenges, there are some related studies. Zhang *et al.* in [5] initially introduced a concept of Perceptive Mobile Network (PMN) and proposed a Remote Radio Units (RRUs) cooperative sensing scheme under the architecture of Centralized Radio Access Network (C-RAN) to improve sensing accuracy. In [6], Ji *et al.* proposed a broad concept of cooperative sensing in ISAC system, including multi-static, multi-band, and multi-source cooperation. Tong *et al.* in [7] further presented the concept, algorithm, and demonstration of multi-view cooperative sensing in ISAC system. References [8, 9] explored the sensing information fusion algorithms with multi-BS cooperative sensing. Zhang *et al.* in [10] proposed multi-node cooperative passive sensing system and a sensing information fusion localization scheme. Overall, existing studies on cooperative sensing in ISAC system mainly focus on multiple BSs cooperative sensing. However, even in terms of node-level cooperative sensing, such as our previous work [9, 14], the cooperation between BS and UE, the cooperation between macro BS and micro BS, and the cooperation between BS and UAV are rarely studied. Besides, the multi-BS cooperation is not structurally classified.

This article aims to provide a deep, comprehensive and concise view on cooperative sensing in ISAC system, namely, the cooperation in the perspectives of resource, node and infrastructure. The main contributions of this article are as follows.

1. In terms of resource-level cooperation, the echo signals in multiple antennas, multiple time slots or frames, multiple frequency bands, and multiple code words could be fused to improve sensing accuracy. Meanwhile, the sensing information in multi-domain resources could be fused simultaneously to further enhance the sensing accuracy.
2. In terms of node-level cooperation, we have provided a concise classification and structural relation for multi-node cooperative sensing. Multiple macro BSs or multiple micro BSs could perform cooperative sensing. Given the proximity to the target and potential deployment on high-frequency bands, micro BSs exhibit higher sensing accuracy compared to macro BSs. Hence, multiple micro BSs cooperative sensing is a preferred scheme. Nevertheless, multiple micro BSs have smaller overlapping areas. When the target moves beyond the overlapped coverage of multiple micro BSs, cooperation between macro and micro BSs becomes necessary. Meanwhile, the cooperation schemes between BS and UE are also studied in this article. Multi-node cooperation effectively improves sensing accuracy and continuity.
3. In terms of infrastructure-level cooperative sensing, the infrastructures from one or multiple operators, across the air, ground and space, could be applied to realize seamless target sensing. The network architecture supporting cooperative sensing, the moving target detection and tracking methods, and the digital twin assisted large-coverage cooperative sensing, are investigated in the infrastructure-level cooperation.

The remainder of this article is organized as follows. The basic concepts in ISAC system are firstly introduced. Then, the resource-level, node-level, and infrastructure-level cooperations are revealed, respectively. Finally, the performance evaluation is provided and the key viewpoints of this article are summarized.

Basic Concepts in ISAC System

In this section, we provide a brief overview of the types of sensing, performance metrics, scenarios, and requirements within the ISAC system.

Types of Sensing

The radar sensing in ISAC system mainly includes active sensing and passive sensing. In active sensing, the transceiver detects the target by receiving the echo signal reflected by the target. In passive sensing, the receiver detects the target by receiving the signal emitted from the target or the echo signal from other transmitters reflected by the target [5]. The BS with sensing function realizes the coupling of digital space and physical space, which is a unified information infrastructure supporting IoE.

Performance Metrics of Sensing

- **Accuracy of detection:** The accuracy of detection is measured by the probabilities of detection and false alarm. This performance metric is primarily applied in the scenario of intrusion detection for UAVs and pedestrians.
- **Accuracy of parameter estimation:** The accuracy of the estimation of distance and velocity of target is measured by Mean Square Error (MSE), Root MSE (RMSE) and Normalized MSE (NMSE), which measure the deviation between the estimation value and the real value. In the scenario of Internet of Vehicles (IoV), this performance metric is used to evaluate the feasibility of obstacle avoidance for vehicles and environmental reconstruction. It is also used to evaluate the effectiveness and superiority of sensing signal processing methods.
- **Sensing area:** Sensing area measures the range of radar sensing, which is influenced by the fading of radar signal and the Radar Cross-Section (RCS) of target. The assessment of sensing area is a prerequisite for the operation of sensing service, which simultaneously facilitates the implementation of multi-node cooperation.
- **Sensing continuity:** Sensing continuity measures the performance of target tracking. When detecting the target in n continuous time instants, the fraction of the time instants

The deep cooperation of ISAC system includes the resource-level cooperation, node-level cooperation, and infrastructure-level cooperation. The resource-level cooperation meanly improves the sensing accuracy. The node-level cooperation meanly improves the sensing area and accuracy. The infrastructure-level cooperation meanly improves the sensing continuity.

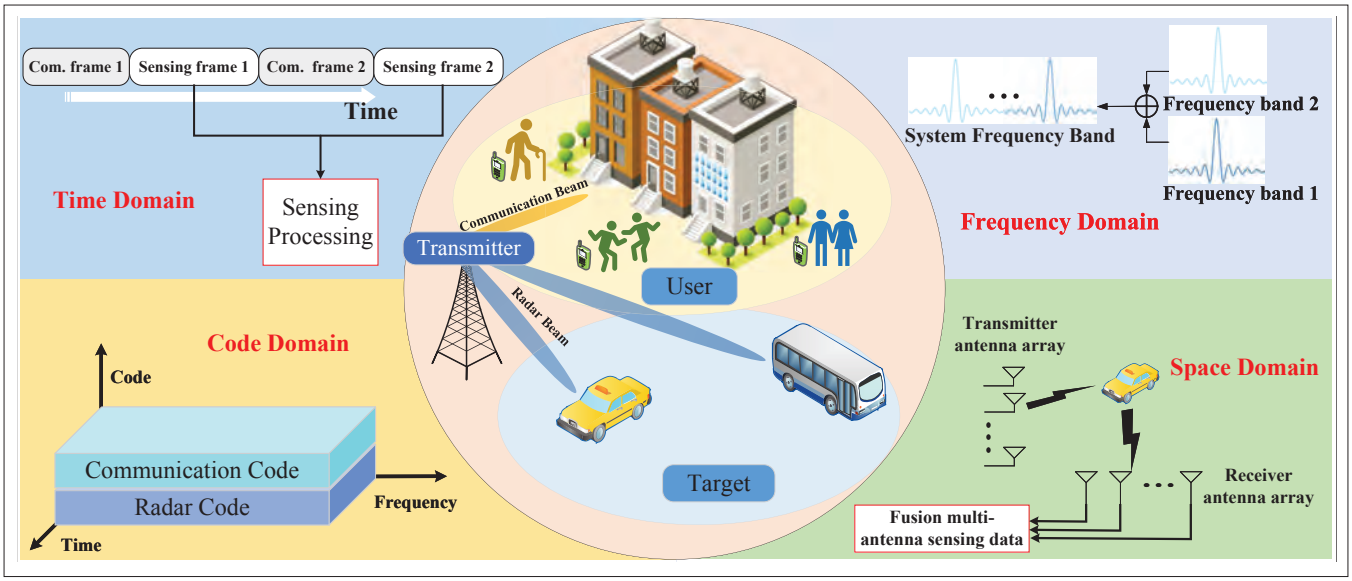


FIGURE 1. Resource-level cooperation in multi-domain resources.

with sensing accuracy higher than a threshold to n is defined as the sensing continuity. Sensing continuity is a performance metric applied in target tracking scenarios.

Scenarios and Requirements

- **Cooperative Sensing of Vehicles:** In the application of intelligent transportation, the lane-level sensing of vehicles needs to be realized. Thus, cooperative sensing is essential in this scenario to improve the sensing accuracy.
- **Cooperative Sensing of UAVs:** In the scenario of UAV sensing in the urban areas, the cooperative sensing is required to realize continuous sensing of UAVs with high maneuverability due to the blockage of buildings. Since the RCS of UAV is small, the cooperative sensing is required to improve the accuracy of small target sensing.
- **Cooperative sensing of other devices:** In the scenario of smart home and smart industry, high-accuracy, comprehensive, and low-latency sensing services are required for the services such as human respiratory detection, illegal intrusion detection, environmental reconstruction in factories, etc. Therefore, cooperative sensing among Wi-Fi, indoor small cells, terminals, and other sensing devices is essential.

Resource-Level Cooperation

As shown in Fig. 1, within a single node, the resource-level cooperation could be realized to improve the accuracy and resolution of sensing, which improves the accuracy of detection and parameter estimation. Resource-level cooperation, as the cooperation in microscopic perspective, could be combined with the other levels of cooperation to further improve the sensing performance.

Cooperation in Space-Domain Resource

Resource cooperation in space-domain mainly applies MIMO to exploit the multi-path signal propagation. As shown in Fig. 1, the signals in multiple antennas could be fused to enhance the accuracy of sensing, which is further classified into the sensing information fusion with uniform and non-uniform antenna array. Meanwhile, multi-antenna beamforming can reduce the interference between communication and sensing.

Sensing Information Fusion over Uniform Antenna Array: In single-node sensing scenario, the angle, distance and velocity of target need to be estimated for localization and trajectory prediction. With a uniform antenna array, the number of antennas affects the resolution of angle estimation for the target, which is realized using Multiple Signal Classification (MUSIC) method, Estimating Signal Parameter via Rotational Invariance Techniques (ESPRIT) method, etc [4]. On the other hand, the accuracy and resolution of distance estimation are related to bandwidth and Signal-to-Noise Ratio (SNR) of echo signal, and the accuracy and resolution of velocity estimation are related to sensing time and SNR. Since multiple antennas occupy the same time-frequency resources, the SNR of echo signal can be improved by fusing the sensing information on multiple antennas, further improving the performance of distance and velocity estimation, which is realized by phase compensation and correlation accumulation. Meanwhile, the time-frequency resources are extended by utilizing the data of multi-antenna, i.e., enlarging the bandwidth and extending sensing time, which is accomplished by splicing the data in each antenna with the challenge of determining the initial phase of the data in each antenna to perform splicing operation.

Sensing Information Fusion Over Non-Uniform Antenna Array: For non-uniform antenna arrays, such as sparse antenna arrays, optimization algorithms can be employed to achieve the performance of a full antenna array using only a portion of antenna resources. For the estimation of target's angle, the virtual antenna array is obtained by hybrid cross-multiplication among antennas, so that the angle can be estimated using MUSIC method. In contrast to uniform antennas, hybrid cross-multiplication leads to a decrease in the SNR on multi-antenna, which reduces the SNR gain in the sensing information fusion of multi-antenna and further degrades the sensing performance. Hence, the trade-off between improving sensing

performance of multi-antenna fusion and saving space-domain resources exists with the non-uniform antenna array.

Cooperation in Time-Domain Resource

The cooperation in time-domain, as shown in Fig. 1, is mainly used to improve the SNR of echo signal by time accumulation, further improving the estimation accuracy of distance and velocity. Assuming that ISAC system adopts OFDM signal, the multiple frames can be used for coherent or non-coherent accumulation to improve the SNR of echo signal and further improve the estimation accuracy of distance. Similarly, multiple frames can be used to improve the resolution of velocity estimation.

Cooperation in Frequency-Domain Resource

Cooperation in frequency-domain mainly involves the fusion of the multiple reference signals and the signals on multiple frequency bands, as shown in Fig. 1.

Fusion Over Multiple Reference Signals: Reference signals, also known as pilot signals, can be applied in radar sensing [11], which include downlink and uplink reference signals. Downlink reference signals could be applied in downlink active or passive sensing. The uplink reference signals including the Demodulation Reference Signal (DMRS), Sounding Reference Signal (SRS), and so on, could be applied in uplink sensing. Since different reference signals occupy different resources in frequency domain, the various reference signals can be cooperatively used in radar sensing, achieving higher sensing performance than single reference signal. Given that multiple reference signals do not fill the complete time-frequency resource blocks, the challenge of this research is the sidelobes deterioration caused by non-continuous resources in time-frequency domains. The sparsity of targets when mapping the data in the time-frequency domain to the Doppler-delay domain brings an opportunity for the application of Compressed Sensing (CS) technique into ISAC signal processing. The CS technique can be applied to recover sensing information at empty position in the time-frequency resource blocks. Therefore, the CS-based multiple reference signals cooperative sensing has significant performance improvement.

Fusion Over High and Low Frequency Bands: The mmWave and THz frequency bands are gradually applied in future mobile communication system. Meanwhile, the sub-6 GHz frequency bands are essential in enhancing the coverage of mobile communication system. With the technique of Carrier Aggregation (CA), the high and low-frequency bands are combined to improve the performance of communication. Similarly, the cooperation of high and low-frequency bands can also improve the sensing performance, where the challenge is the inconsistency of physical-layer parameters in the high and low-frequency bands during sensing information fusion. For example, when adopting OFDM as the ISAC signal, the different subcarrier spacings in the high and low-frequency bands brings challenge for sensing information fusion. This problem can be solved by reorganizing the channel information matrices of high and low-frequency bands to match the parameters of the corresponding positional elements [12]. In addition to the cooperation of high and low-frequency bands, there are additional cases of cooperation in frequency-domain, such as the fusion of the signals over the fragmented spectrum bands or unlicensed frequency bands.

Cooperation in Code-Domain Resource

With the cooperation in code-domain, communication and radar sensing functions share the same resources in time-frequency domains, distinguished by code words. This approach enables higher resolution of distance and velocity estimation [13], as the resources in time-frequency domains are not diminishing with the code-division method compared with the traditional time-division or frequency-division method. Furthermore, the sensing information on different code words within the same time-frequency domains can be extracted and fused to obtain high sensing accuracy.

Cooperation in Multi-Domain Resources

In mobile communication systems, it is common to enhance communication performance through the utilization of multi-domain resources, which makes the multi-domain cooperative sensing feasible. The cooperative sensing in multi-domain fuses the sensing information in multi-domain resources to achieve a higher sensing accuracy than the sensing method using single-domain resource. For example, using multiple antennas, multiple signals on different antennas may overlap in the time-frequency domains, so that different code words are applied to distinguish the multiple signals. Then, the multiple signals with different code words could be extracted and fused, thereby improving the sensing accuracy.

In urban environments, the sensing information in multi-path could be applied to improve the sensing accuracy with multi-domain resource cooperation. The challenge of multi-path research is multi-path separation. Common methods include blind source separation and independent component analysis. However, the features of different paths are generally varied in multi-domain. Therefore, the multi-path mixed signals can be separated through multi-domain resource cooperation, which are further fused to realize high-precision target localization, high-continuity target tracking, and high-accuracy environmental reconstruction.

Node-Level Cooperation

In node-level cooperation, multiple nodes, including macro BS, micro BS, UE, could cooperate to improve the accuracy of detection, the accuracy of parameter estimation, and the sensing area mentioned earlier. Besides, node-level cooperation can be combined with resource-level cooperation to further improve the sensing performance. However, compared with resource-level cooperation, the node-level cooperation obtains the capability of multi-view sensing, which has potential in Three-Dimensional (3D) imaging for the target and environment. Therefore, compared to resource-level cooperation, node-level cooperation is a cooperation in mesoscopic perspective.

Multi-BS Cooperation

The cooperation between multiple BSs is categorized according to the type of BS, namely multiple micro BSs cooperative sensing, macro-micro

The resource-level cooperation includes space-domain cooperation, time-domain cooperation, frequency-domain cooperation, code-domain cooperation, and multi-domain cooperation. Especially, the cooperation in multi-domain further improves the sensing accuracy.

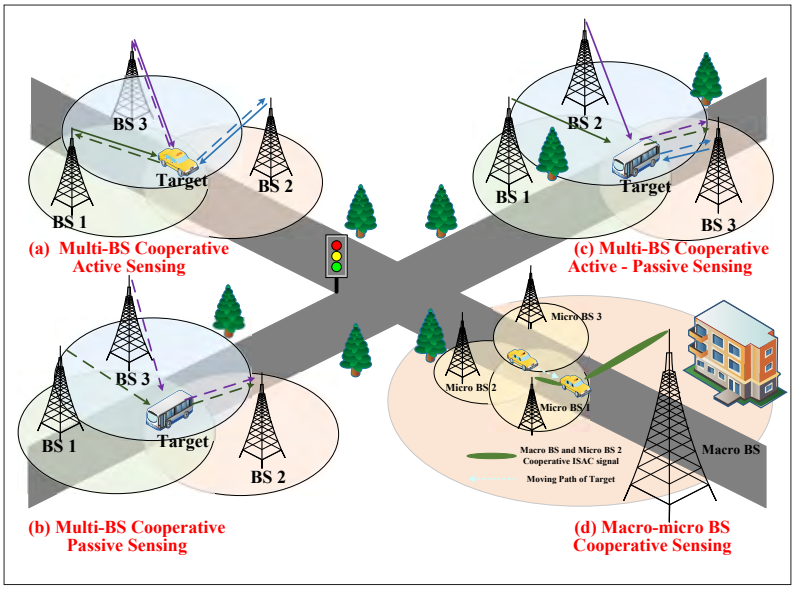


FIGURE 2. Multi-BS cooperation.

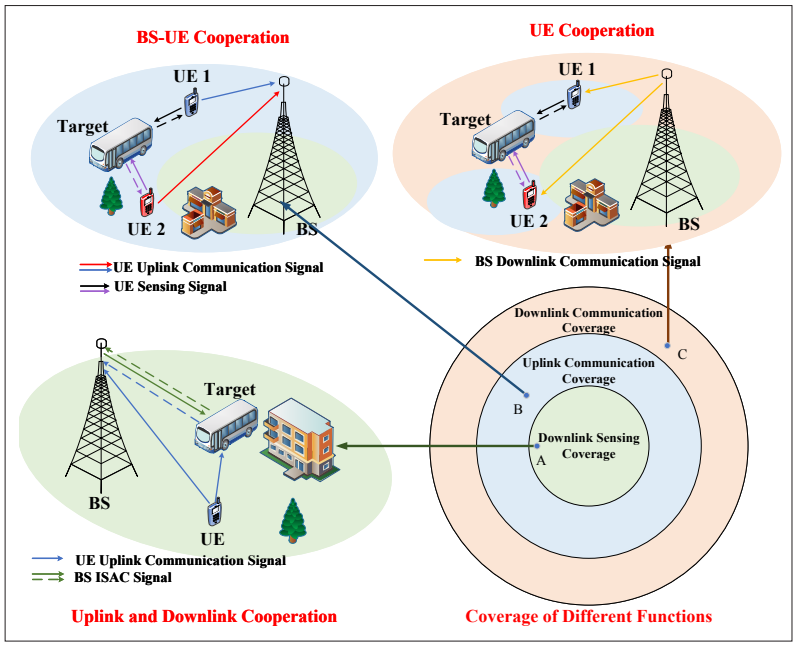


FIGURE 3. BS-UE cooperation.

BSs cooperative sensing, and multiple macro BSs cooperative sensing, where the sensing area is increasing in these three categories.

Figure 2a shows multi-BS cooperative active sensing, i.e., when the target is located in the overlapped coverage area of multiple BSs, each BS performs sensing separately, and the sensing information fusion is performed in the fusion center, which could be one BS or the Mobile Edge Computing Server (MECS) [14]. The challenges of cooperative active sensing lie in spatial registration and symbol-level fusion of multiple non-coherent echo sensing information [9, 14]. Figure 2b shows multi-BS cooperative passive sensing, i.e., the passive BS receives the echo signals of other BSs reflected by the target and performs sensing information fusion. The challenges of cooperative passive sensing are how to eliminate the synchronization error and achieve a symbol-level fusion sensing [9, 14]. Figure 2c shows multi-BS cooperative active-passive sensing, i.e., the BS can both perform active sensing and receive the echo signals of other BSs reflected by the target in passive sensing, and finally fuse the sensing information of active sensing and passive sensing. The challenges of cooperative active sensing and cooperative passive sensing both exist in this type of cooperation. The above three types of cooperative sensing requires that the target is located in the overlapped coverage area of multiple BSs. The future research trends in multi-BS cooperation include multi-BS space-time-frequency synchronization, multi-BS mutual interference mitigation, multi-BS cooperative beamforming and target tracking, multi-BS sensing information fusion. In addition, since each BS can independently perform resource-level cooperation, the resource allocation in resource-level cooperation is also the future research trend. Since the sensing area of micro BS is smaller than that of macro BS, when the target moves out of the sensing area of micro BS, the macro-micro BSs cooperative sensing is performed, as shown in Fig. 2d, which is investigated next.

Macro-micro BSs Cooperation

Compared with macro BS, micro BS has higher frequency band and smaller transmit power. Hence, the sensing area of micro BS is smaller than that of macro BS. When the target moves out of the overlapped sensing area of multiple micro BSs as shown in Fig. 2d, the macro BS and micro BS could cooperate to improve the sensing accuracy and continuity. The type of macro-micro BSs cooperation also consists of cooperative active, passive, and active-passive sensing. Macro-micro BSs cooperation not only encounters the challenges similar to those in multi-BS cooperation but also faces the challenge of the fusion of the sensing information from the micro BS working on high-frequency band and the macro BS working on low frequency band. To address the above challenge, the channel information matrices from macro and micro BSs are adjusted and fused with low synchronization accuracy, which achieves higher sensing accuracy compared with data-level sensing information fusion.

BS-UE Cooperation

The BS and UE cooperative sensing can be classified according to the sensing range. Generally, the downlink sensing range is smaller than the uplink communication range, and the uplink communication range is smaller than the downlink communication range, as shown in Fig. 3. When the target is located within the downlink sensing coverage, such as the location A in Fig. 3, the BS can simultaneously perform downlink sensing by receiving the echo signal of BS and uplink sensing by receiving the uplink reflected signal from UE. Then, the echo signals from downlink and uplink sensing are fused in the BS, which is the cooperative downlink and uplink sensing. When the target is outside of the downlink sensing coverage and within the uplink communication coverage, such as the location B in Fig. 3, multiple UEs detect the target and upload the sensing information to the BS for sensing information fusion. When the target is outside of the uplink communication coverage and within the downlink communication

coverage, such as the location C in Fig. 3, multiple UEs detect the target and fuse the sensing information by themselves, with the guidance of BS in resource allocation. The challenge of BS and UE cooperative sensing is to mitigate the uplink and downlink interference, achieving the mutual benefits of sensing and communication. Hence, redesigning the frame structure is essential, coupled with leveraging uplink sensing information for beamforming.

Infrastructure-Level Cooperation

In the application of smart city, there are a large number of targets to be detected and tracked. For example, in the scenario of UAV sensing, due to the high maneuverability of UAV, the infrastructure-level cooperation is required to accurately detect and track UAV. Infrastructure-level cooperation is expected to mainly enhance the sensing continuity, which achieves wider area sensing compared with the node-level and resource-level cooperation, representing the cooperation in macroscopic perspective. Generally, the following techniques are required in the infrastructure-level cooperation.

Network Architecture Supporting Cooperative Sensing: In order to support ISAC enabled cooperative sensing, the network architecture needs to be designed. The network elements include Remote Radio Unit (RRU), Building Base Band Unite (BBU), MECS, and core network, etc. The sensing information fusion centers are deployed in the MECS for node-level cooperative sensing and deployed in the core network for infrastructure-level cooperative sensing. Multiple RRUs detect target with multi-domain resources and fuse the sensing information in the MECS. The interfaces between BSs, information processing procedures, and signaling interaction procedures among the network elements need to be designed to support cooperative sensing. The fusion center in the core network fuse the large amount of sensing information to build a full view of the digital twin of physical space.

In order to realize large-scale and seamless sensing, the Heterogeneous Networks (HetNets), including the BSs from different operators and the Access Points (APs) using IEEE 802 techniques, could cooperate to sense target and environment. In this case, the network architecture supporting HetNets cooperative sensing needs to be designed. A new framework for fusing sensing information from HetNets needs to be constructed to support network interoperability. Then, the procedure of sensing information fusion with non-synchronization among HetNets faces great challenges. Hence, the space and time calibration among HetNets is necessary.

Moving Target Detection and Tracking: When tracking the target with high maneuverability, the infrastructures from the space-air-ground integrated networks, as well as the infrastructures from single or multiple operators, could cooperatively detect and track the target. As shown in Fig. 4, multi-BS cooperation, as well as the cooperation between BS and UAV, could be performed to detect the vehicle and UAV with high maneuverability. In this scenario, the handover of target sensing by multiple BSs needs to be studied. The handover in target sensing studies the switching of multiple BSs belonging to the same operator or different operators to realize the continuous sensing of moving target due to the limited coverage of single BS. Besides, the detection methods of moving small target, and the machine learning techniques for target tracking need to be studied.

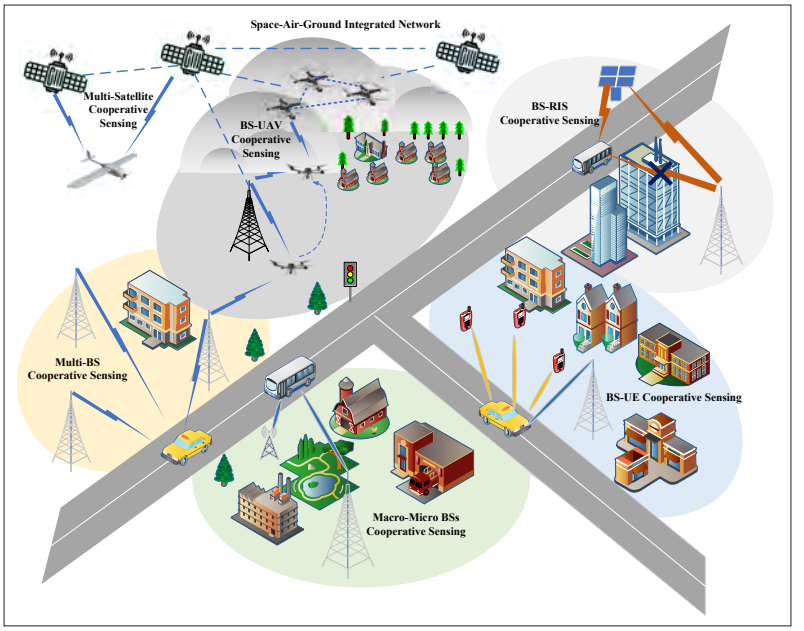


FIGURE 4. Infrastructure-level cooperation.

Digital Twin Assisted Infrastructure-Level Cooperation: In the scenario of intelligent transportation, the massive amount of sensing information needs to be structurally stored and processed. Furthermore, heterogeneous infrastructures bring diverse protocols and processing methods for sensing information. Therefore, the storage, management and fusion of sensing information are difficult in the infrastructure-level cooperation. Digital twin can not only replace tedious and inefficient manual management of sensing information, achieving low-cost, visual, and intelligent management of sensing information, but also enables optimal distributed resources scheduling. Digital twin is a potential solution for massive sensing information fusion in the infrastructure-level cooperation [15].

In infrastructure-level cooperation, the construction and updating of digital twin face some challenges. On the one hand, the large number of devices bring great difficulties for the low-delay construction and updating of digital twin entities. In this case, one of the viable solutions is to construct and update the digital twin entities in a distributed manner. On the other hand, given the high mobility of vehicles and UAVs, as well as the frequent switching of different devices, ensuring the low-delay updating of the sensing information and the continuity of sensing is necessary.

Performance Evaluation

In this section, we provide the simulation results of resource-level and node-level cooperative sensing, verifying the advantages of cooperative sensing. Among the performance metrics mentioned earlier, the performance metric of RMSE is commonly used to characterize sensing accuracy.

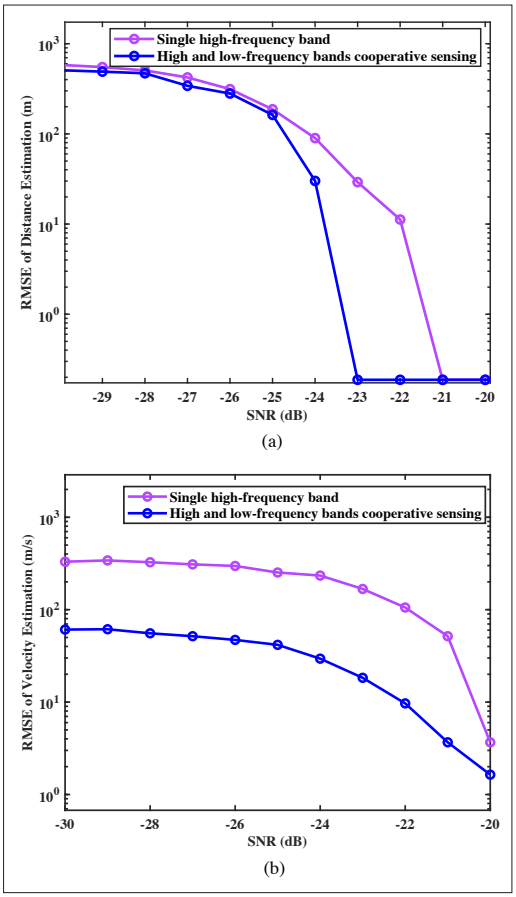


FIGURE 5. RMSEs of multi-band sensing and single-band sensing: a) range estimation; b) velocity estimation.

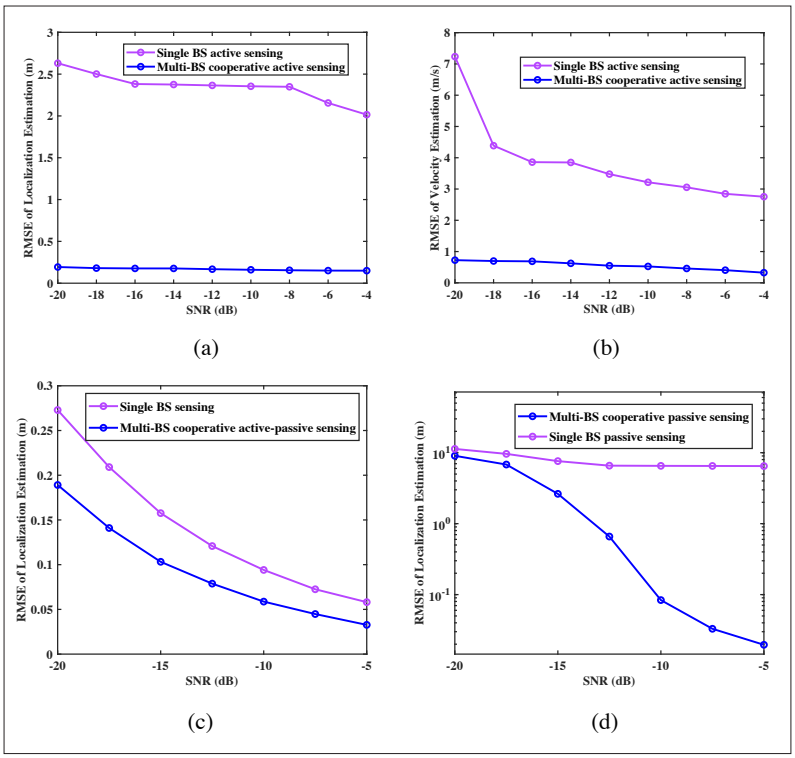


FIGURE 6. RMSEs of multi-BSs cooperative sensing: a) target localization with multi-BS cooperative active sensing; b) velocity estimation with multi-BS cooperative active sensing; c) target localization with multi-BS cooperative active-passive sensing; d) target localization with multi-BS cooperative passive sensing.

Resource-Level Cooperation

According to the 3rd Generation Partnership Project (3GPP) 38.211 standard, the high-frequency band with carrier frequency of 24 GHz and subcarrier spacing of 120 kHz, and the low-frequency band with carrier frequency of 5.9 GHz and subcarrier spacing of 30 kHz, are applied in the high and low frequency bands cooperative sensing [12]. As illustrated in Fig. 5, it is obvious that the RMSEs of distance and velocity estimation are lower and the convergence speed is faster with multiple frequency bands cooperative sensing than those results with single frequency band sensing. The reason is that the time-frequency resources are extended with multiple frequency bands cooperation.

Node-Level Cooperation

As depicted in Fig. 6, the multi-BS cooperative UAV sensing is simulated to verify the performance of node-level cooperative sensing. The number of BSs is 4, the subcarrier spacing is 60 kHz, and the carrier frequency is 4.9 GHz. Multi-BS cooperative active sensing is compared with single-BS sensing in terms of the performance of localization and velocity estimation for UAV. As illustrated in Figs. 6a and 6b, the multi-BS cooperative sensing achieves more accurate localization estimation than single-BS sensing. Multi-BS cooperative active sensing achieves significantly better performance in velocity estimation than single-BS sensing. The same simulation parameters are applied in multi-BS cooperative active-passive sensing and multi-BS cooperative passive sensing, where 1000 times Monte Carlo simulations are performed to calculate the RMSE of location estimation. It is revealed that the localization performance is significantly better with multi-BS cooperative active-passive sensing, as well as with multi-BS cooperative passive sensing, compared to single-BS sensing, as shown in Figs. 6c and 6d.

Conclusion

In order to realize high-accuracy, large-coverage, and continuous sensing in ISAC system, this article provides a deep and comprehensive view on the cooperative sensing in ISAC system, including resource-level cooperative sensing, node-level cooperative sensing, and infrastructure-level cooperative sensing. In the resource-level cooperation, the sensing information in time-frequency-space-code domains is fused to improve sensing accuracy. In node-level cooperation, multiple nodes, including BS and UE, could perform cooperative sensing to fuse the sensing information from multiple nodes, extending the sensing coverage and improving the sensing accuracy. In infrastructure-level cooperation, the large number of infrastructures perform cooperative sensing to realize continuous sensing. The research in this article may provide a research guideline for cooperative sensing in ISAC system, promoting the applications of IoE with the connection of digital and physical spaces.

References

[1] F. Liu *et al.*, "Integrated Sensing and Communications: Toward Dual- Functional Wireless Networks for 6G and Beyond," *IEEE JSAC*, vol. 40, no. 6, 2022, pp. 1728–67.
[2] ITU, "Future Technology Trends of Terrestrial International Mobile Telecommunications Systems Towards 2030 and Beyond," 2022.

[3] F. Liu *et al.*, "Joint Radar and Communication Design: Applications, State-of-the-Art, and the Road Ahead," *IEEE Trans. Commun.*, vol. 68, no. 6, June 2020, pp. 3834–62.
[4] Z. Wei *et al.*, "Integrated Sensing and Communication Signals Toward 5G-A and 6G: A Survey," *IEEE Internet of Things J.*, vol. 10, no. 13, 2023, pp. 11,068–92.
[5] A. Zhang *et al.*, "Perceptive Mobile Networks: Cellular Networks With Radio Vision via Joint Communication and Radar Sensing," *IEEE Vehic. Tech. Mag.*, vol. 16, no. 2, June 2021, pp. 20–30.
[6] K. Ji *et al.*, "Networking Based ISAC Hardware Testbed and Performance Evaluation," *IEEE Commun. Mag.*, vol. 61, no. 5, May 2023, pp. 76–82.
[7] X. Tong, Z. Zhang, and Z. Yang, "Multi-View Sensing for Wireless Communications: Architectures, Designs, and Opportunities," *IEEE Commun. Mag.*, vol. 61, no. 5, May 2023, pp. 40–46.
[8] G. Li *et al.*, "Multi-Point Integrated Sensing and Communication: Fusion Model and Functionality Selection," *IEEE Wireless Commun. Letters*, vol. 11, no. 12, Dec 2022, pp. 2660–64.
[9] Z. Wei *et al.*, "Symbol-Level Integrated Sensing and Communication Enabled Multiple Base Stations Cooperative Sensing," *IEEE Trans. Vehic. Tech.*, Aug 2023, pp. 1–15.
[10] Z. Zhang *et al.*, "Target Localization and Performance Trade-Offs in Cooperative ISAC Systems: A Scheme Based on 5G NR OFDM Signals," arXiv preprint arXiv:2403.02028, 2024.
[11] Z. Wei *et al.*, "Multiple Reference Signals Collaborative Sensing for Integrated Sensing and Communication System Towards 5G-A and 6G," *IEEE Trans. Vehic. Tech.*, 2024, pp. 1–15.
[12] Z. Wei *et al.*, "Carrier Aggregation Enabled Integrated Sensing and Communication Signal Design and Processing," *IEEE Trans. Vehic. Tech.*, Oct 2023.
[13] X. Liu *et al.*, "Complementary Coded Scrambling RadCom System–An Integrated Radar and Communication Design in Multi-User-Multi-Target Scenarios," *IEEE Trans. Vehic. Tech.*, 2023.
[14] Z. Wei *et al.*, "Integrated Sensing and Communication Enabled Multiple Base Stations Cooperative Sensing Towards 6G," *IEEE Network*, 2023.
[15] Z. Wei *et al.*, "Integrated Sensing and Communication Driven Digital Twin for Intelligent Machine Network," accepted for publication in *IEEE Internet of Things Mag.*, 2024.

Biographies

Zhiqing Wei [M] (weizhiqing@bupt.edu.cn) received the B.E. and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2010 and 2015, respectively. He is an Associate Professor with BUPT. He has authored one book, three book chapters, and more than 50 papers. His research interest is the performance analysis and optimization of intelligent machine networks. He was granted the Exemplary Reviewer of *IEEE Wireless Communications Letters* in 2017, the Best Paper Award of WCSP 2018. He was the Registration Co-Chair of IEEE/CIC ICCC 2018, the publication Co-Chair of IEEE/CIC ICCC 2019 and IEEE/CIC ICCC 2020.

Haotian Liu [StM] (haotian_liu@bupt.edu.cn) received the B.E. degree in School of Physic and Electronic Information Engineering, Henan Polytechnic University (HPU) in 2023. He is currently pursuing his M.S. degree with Beijing University of Posts and Telecommunication (BUPT). His research interests include integrated sensing and communication, cooperative sensing, compressed sensing, carrier aggregation.

Zhiyong Feng [M'08, SM'15] (fengzy@bupt.edu.cn) received her B.E., M.E., and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), Beijing, China. She is a professor at BUPT, and the director of the Key Laboratory of the Universal Wireless Communications, Ministry of Education, P.R.China. She is a senior member of IEEE, vice chair of the Information and Communication Test Committee of the Chinese Institute of Communications (CIC). Currently, she is serving as Associate Editors-in-Chief for China Communications, and she is a technological advisor for international forum on NGMN. Her main research interests include wireless network architecture design and radio resource management in 5th generation mobile networks (5G), spectrum sensing and dynamic spectrum management in cognitive wireless networks, and universal signal detection and identification.

Huici Wu [M] (dailywu@bupt.edu.cn) received the Ph.D degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2018. From 2016 to 2017, she visited the Broadband Communications Research (BBCR) Group, University of Waterloo, Waterloo, ON, Canada. She is now an Associate Professor at BUPT. Her research interests are in the area of wireless communications and networks, with current emphasis on collaborative air-to-ground communication and wireless access security.

Fan Liu [SM] (liuf6@sustech.edu.cn) received the B.Eng. and Ph.D. degrees from Beijing Institute of Technology (BIT), Beijing, China, in 2013 and 2018 respectively, respectively. He is currently an Assistant Professor with the School of System Design and Intelligent Manufacturing (SDIM), Southern University of Science and Technology (SUSTech). He has previously held academic positions with the University College London (UCL), first as a Visiting Researcher from 2016 to 2018, and then as a Marie Curie Research Fellow from 2018 to 2020. His research interests include signal processing and wireless communications, and in particular in the area of Integrated Sensing and Communications (ISAC).

Qixun Zhang [M] (zhangqixun@bupt.edu.cn) received the B.E. and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2006 and 2011, respectively. From March 2006 to June 2006, he was a Visiting Scholar with the University of Maryland, College Park, MD, USA. From November 2018 to November 2019, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA. He is currently a Professor with the Key Laboratory of Universal Wireless Communications, Ministry of Education, School of Information and Communication Engineering, BUPT. His research interests include 5G mobile communication systems, integrated sensing and communication for autonomous driving vehicle, mmWave communication systems, and unmanned aerial vehicles (UAVs) communication.

Yucong Du [StM] (duyc@bupt.edu.cn) received the B.S. degree and the M.S. degree from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2020 and 2023. He is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His current research interests include integrated sensing and communication network, digital twin network and resource management.

IEEE Access®

Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Harness the publishing power of IEEE *Access*®.

IEEE *Access* is a multidisciplinary open access journal offering high-quality peer review, with an expedited, binary review process of 4 to 6 weeks. As a journal published by IEEE, IEEE *Access* offers a trusted solution for authors like you to gain maximum exposure for your important research.

Explore the many benefits of IEEE *Access*:

- **Receive high-quality, rigorous peer review** in only 4 to 6 weeks
- **Reach millions of global users** through the IEEE *Xplore*® digital library by publishing open access
- **Submit multidisciplinary articles** that may not fit in narrowly focused journals
- **Obtain detailed feedback** on your research from highly experienced editors
- **Establish yourself as an industry pioneer** by contributing to trending, interdisciplinary topics in one of the many topical sections IEEE *Access* hosts
- **Present your research to the world quickly** since technological advancement is ever-changing
- **Take advantage of features** such as multimedia integration, usage and citation tracking, and more
- **Publish without a page limit** for $1,750 per article

⋮ Learn more at ieeeaccess.ieee.org

IEEE